Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation

Kuzmanovic, Bojana ; Rigoux, Lionel ; Tittgemeyer, Marc

Abstract: When updating beliefs about their future prospects, people tend to disregard bad news. By combining fMRI with computational and dynamic causal modeling, we identified neurocircuitry mechanisms underlying this optimism bias to test for valence-guided belief formation. In each trial of the fMRI task, participants (n = 24, 10 male) estimated the base rate (eBR) and their risks of experiencing negative future events, were confronted with the actual BR, and finally had the opportunity to update their initial self-related risk estimate. We demonstrated an optimism bias by revealing greater belief updates in response to good over bad news (i.e., learning that the actual BR is lower or higher than expected) while controlling for confounds (estimation error and personal relevance of the new information). Updating was favorable when the final belief about risks improved (or at least did not worsen) relative to the initial risk estimate. This valence of updating was encoded by the ventromedial prefrontal cortex (vmPFC) associated with the valuation of rewards. Within the updating circuit, the vmPFC filtered the incoming signal in a valence-dependent manner and influenced the dorsomedial prefrontal cortex (dmPFC). Both the valence-encoding activity in the vmPFC and its influence on the dmPFC predicted individual magnitudes of the optimism bias. Our results indicate that updating was biased by the motivation to maximize desirable beliefs, mediated by the influence of the valuation system on further cognitive processing. Therefore, although it provides the very basis for human reasoning, belief formation is essentially distorted to promote desired conclusions.SIGNIFICANCE STATEMENT The question of whether human reasoning is biased by desires and goals is crucial for everyday social, professional, and economic decisions. How much our belief formation is influenced by what we want to believe is, however, still debated. Our study confirms that belief updates are indeed optimistically biased. Critically, the bias depends on the recruitment of the brain valuation system and the influence of this system on neural regions involved in reasoning. These neurocircuit interactions support the notion that the motivation to maximize pleasant beliefs reinforces those cognitive processes that are most likely to yield the desired conclusion.

DOI: https://doi.org/10.1523/jneurosci.0266-18.2018

JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation

Bojana Kuzmanovic[1], Lionel Rigoux[1,2] and Marc Tittgemeyer[1]

[1]*Translational Neurocircuitry Group, Max Planck Institute for Metabolism Research Cologne, 50931 Cologne, Germany*
[2]*Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, 8032 Zurich, Switzerland*

DOI: 10.1523/JNEUROSCI.0266-18.2018

Received: 30 January 2018

Revised: 17 July 2018

Accepted: 20 July 2018

Published: 13 August 2018

Author contributions: B.K. designed research; B.K. performed research; B.K. and L.R. analyzed data; B.K. wrote the first draft of the paper; B.K., L.R., and M.T. edited the paper; B.K., L.R., and M.T. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

We thank Morné Truter for proofreading and valuable comments on an earlier draft of the manuscript and Thorben Huelsduenker for assistance with data collection.

Corresponding Author: Bojana Kuzmanovic, Max Planck Institute for Metabolism Research, Translational Neurocircuitry Group, Gleuelerstr. 50, 50931 Cologne, Germany, Tel.: +49 221 4726-255, Email: bojana.kuzmanovic@sf.mpg.de

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.0266-18.2018

Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation

Abbreviated Title:

Neurocircuits of Valence-Guided Belief Formation

Bojana Kuzmanovic[1*], Lionel Rigoux[1,2], Marc Tittgemeyer[1]

[1] Translational Neurocircuitry Group, Max Planck Institute for Metabolism Research Cologne, 50931 Cologne, Germany

[2] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, 8032 Zurich, Switzerland

* Corresponding Author:
Bojana Kuzmanovic
Max Planck Institute for Metabolism Research
Translational Neurocircuitry Group
Gleuelerstr. 50, 50931 Cologne, Germany
Tel.: +49 221 4726-255
Email: bojana.kuzmanovic@sf.mpg.de

Number of pages: 51; Number of figures: 4; Number of tables: 4.
Number of words: Abstract 248, Introduction 644, and Discussion 1500.

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgments: We thank Morné Truter for proofreading and valuable comments on an earlier draft of the manuscript and Thorben Huelsduenker for assistance with data collection.

Abstract

When updating beliefs about their future prospects, people tend to disregard bad news. By combining fMRI with computational and dynamic causal modeling, we identified neurocircuitry mechanisms underlying this optimism bias to test for valence-guided belief formation. In each trial of the fMRI task, participants (n = 24, 10 male) estimated the base rate and their risks of experiencing negative future events, were confronted with the actual base rate, and finally had the opportunity to update their initial self-related risk estimate. We demonstrated an optimism bias by revealing greater belief updates in response to good over bad news (i.e., learning that the actual base rate is *lower* or *higher* than expected), while controlling for confounds (estimation error and personal relevance of the new information). Updating was favorable when the final belief about risks *improved* (or at least did not worsened) relative to the initial risk estimate. This valence of updating was encoded by the ventromedial prefrontal cortex (vmPFC) associated with the valuation of rewards. Within the updating circuit, the vmPFC filtered the incoming signal in a valence-dependent manner and influenced the dorsomedial prefrontal cortex (dmPFC). Both the valence-encoding activity in the vmPFC and its influence on the dmPFC predicted individual magnitudes of the optimism bias. Our results indicate that updating was biased by the motivation to maximize desirable beliefs, mediated by the influence of the valuation system on further cognitive processing. Thus, while providing the very basis for human reasoning, belief formation is essentially distorted to promote desired conclusions.

Significance Statement

The question whether human reasoning is biased by desires and goals is crucial for everyday social, professional, and economic decisions. How much our belief formation is influenced by what we *want* to believe is, however, still debated. Our study confirms that belief updates are indeed optimistically biased. Critically, the bias depends on the recruitment of the brain valuation system and the influence of this system on neural regions involved in reasoning. These neurocircuit interactions support the notion that the motivation to maximize pleasant beliefs reinforces those cognitive processes that are most likely to yield the desired conclusion.

58 **Introduction**

59

60 Not only extrinsic rewards such as tasty food, but also internal processes such as
61 desirable beliefs or positive emotions are expected to evoke pleasant states. Believing that one
62 is attractive and intelligent (Eil and Rao, 2011; Korn et al., 2012), or that one's future will be
63 bright (Sharot et al., 2011), has a positive subjective value, which is why people tend to be
64 motivated to maintain such beliefs (Sharot and Garrett, 2016a). In turn, the motivation to
65 maximize pleasant beliefs has been hypothesized to reinforce those cognitive processes that
66 are most likely to yield a desired conclusion (Kunda, 1990; Hughes and Zaki, 2015).
67 However, motivational influences on reasoning have been controversially debated
68 (Kunda, 1990; Shah et al., 2016; Sharot and Garrett, 2016b; Kuzmanovic and Rigoux, 2017).
69 But how can we prove whether specific conclusions are reinforced by desires when these
70 processes are hidden from direct observation and can operate outside of awareness (Tesser,
71 2000)? One way is to identify systematic, valence-dependent biases in information
72 integration. For instance, when participants were given new information relevant to their
73 current belief, they were more likely to incorporate good than bad news (e.g., indicating a
74 *lower* versus *higher* risk than initially expected) to update their belief (Sharot et al., 2011;
75 Kuzmanovic et al., 2016a). Circumventing self-report, such asymmetric updating provides an
76 individual index of the optimism bias by exploiting actual belief formation behavior.
77 Another difficulty is that no reward has a fix subjective value, neither the extrinsic nor
78 the intrinsic ones. Tasty food, for instance, is more pleasant during a hungry state. This is
79 directly reflected in the activity of brain regions encoding the reward value such as the
80 ventromedial prefrontal cortex (vmPFC; Bartra et al., 2013; Chase et al., 2015): neurons in the
81 vmPFC that fired in response to tasty food during a hungry state, no longer showed this
82 response after satiety (Grabenhorst and Rolls, 2011). Likewise, desirable beliefs may be more
83 pleasant after threatening a person's self-worth (Roese and Olson, 2007; Rudman et al.,

2007). Indeed, the magnitude of the optimism bias substantially varied across the participants (e.g., Sharot et al., 2011; Kuzmanovic and Rigoux, 2017). We assume that the belief updating should be biased only in the subjects who indeed assign a positive value to avoiding threatening and enhancing desirable beliefs. This allows us to infer the current value of a specific reward (favorable beliefs) from the reinforcement of the behavior leading to this reward (updating biased towards favorable beliefs).

The present study aims to demonstrate that desirable beliefs have an incentive salience and therefore can guide updates by influencing ongoing cognitive processing. To this end, neural circuits of belief updating were identified by using an established fMRI paradigm. Recently (Kuzmanovic et al., 2016a), we have shown that the vmPFC encoded the positive value of favorable self-related (but nor other-related) belief updates, indicating that the brain transforms beliefs into the same common value scale as classical rewards. Further, we isolated cognitive components and formally controlled for confounds using computational modeling to validate conclusions about valence-dependent update behavior (Kuzmanovic and Rigoux, 2017). Based on this previous work and the central role of the vmPFC in value encoding (Bartra et al., 2013; Chase et al., 2015), we hypothesized that the positive value of favorable updating would be encoded by the vmPFC. Moreover, we expected that only those individuals, who showed an optimism bias, will have a strong neural response to favorable updating. Finally, we used dynamic causal modeling to investigate mechanisms underlying valence-dependent updating. We hypothesized that the contexts of favorable and unfavorable updating would modulate coupling amongst regions recruited during updating, and that the identified valuation system would influence other regions involved in belief formation. Our results provide evidence for a motivationally biased belief formation that is mediated by the value encoding in the vmPFC, and the influence of the vmPFC on the dorsomedial prefrontal cortex.

Materials and Methods

Participants

Forty subjects were recruited from the institute's subject database. A total of four participants was excluded because of problems with task performance. Two participants recognized that the base rates were manipulated, and one individual did not update estimates in 84% of trials (mean of the included sample = 30.09%, SD = 16.13; exclusion threshold = 66.67%). Lastly, one participant updated estimates away from the presented base rate in 18% of trials indicating problems with task understanding (mean of the included sample = 2.66%, SD = 4.12; exclusion threshold = 15%). Data from another 12 participants were excluded due to excessive head movement in the MR scanner that exceeded a threshold of 1.5 framewise displacement (Power et al., 2012). This was necessary to account for increased sensitivity to motion-related artifacts in multiband acquisition for functional imaging (see below for acquisition parameters). Thus, in total 24 subjects were included into the analysis (10 male, mean age = 27.38, *SD* = 5.15). Notably, adding the 12 subjects with excessive motion to the sample of 24 subjects revealed the same behavioral results (see Results, Task Performance).

Experimental Design

The experiment was conducted during the acquisition of fMRI scans using Presentation 18.1 (Neurobehavioral Systems) and consisted of 80 trials with 80 different adverse life events (e.g., cancer or car theft). Participants began each trial with estimating the base rate of an adverse life event (eBR, see **Figure 1**). Next, they were asked to estimate their own likelihood of experiencing the life event in their lifetime (first estimate, E1), and were subsequently presented with the actual base rate (BR). Subjects were instructed that the base rate refers to

the probability of the respective event occurring to persons of the same sex and age, living in the same socio-cultural environment, as determined by the German Federal Statistical Office ("Statistisches Bundesamt"). At the end of each trial, participants had to re-estimate their own risk (second estimate, E2).

The critical behavioral measure was the size of the update, i.e., the difference between E1 and E2. Subjects were expected to update their first risk estimate after being confronted with a base rate different than the one they initially assumed. This difference between eBR and BR indicated the estimation error (EE; EE = |eBR – BR|). In half of the trials, BR was desirable (better than expected, i.e., eBR > BR; good news, GOOD), and in the other half, BR was undesirable (worse than expected, i.e., eBR < BR; bad news, BAD). Notably, we expected participants to change their risk estimates on average toward the new information. That is, upon an actual base rate that is lower than expected, participants should decrease their risk estimates. Conversely, upon an actual base rate that is higher than expected, participants should increase their risk estimates. Indeed, updates toward the direction opposed to the new information were very rare (M = 2.66%, SD = 4.12). This is also reflected in the desirability-dependent computation of updates that ensures that positive values indicate an update toward the new information equally for GOOD and BAD (see Table 1). Valence-dependent bias in updating was present when GOOD and BAD trials yielded different updates (i.e., mean $update_{GOOD}$ > mean $update_{BAD}$ indicates an optimism bias).

Participants were free to report a probability anywhere between 1 and 99%. Starting from 50% in eBR, they selected the desired probability by using two buttons to increase or decrease the number displayed on the screen (**Figure 1**, green font in eBR, E1, and E2), and a third button to confirm the selected choice. Subjects were instructed to use both hands. In the first half of the experiment they used the right hand for selecting the percentage number and

159 the left hand for confirming it, and in the second half the other way around (order counter-bal-
160 anced across subjects). In E1, the starting number equaled the one selected in eBR, and in E2,
161 the starting number corresponded to the one selected in E1.

162 For eBR, E1, and E2, the response display was activated after a 2 s interval. Subjects
163 were instructed to use the first 2 s to think about their estimate, and then had a maximum of
164 10 s to respond (see **Table 1** for mean RTs). BR was presented for 2 s. The intervals within
165 and between the trials consisted of a fixation cross and were jittered (Mumford et al., 2015):
166 the three inter-stimulus intervals within the trial (between eBR and E1, E1 and BR, and BR
167 and E2) ranged between 2375 ms and 4625 ms, with a mean of 3500 ms, and the inter-trial in-
168 tervals ranged between 4875 ms and 7125 ms, with a mean of 6000 ms. The average task du-
169 ration was 48 min (*SD* = 2.45).

170 GOOD and BAD trials were rendered comparable with respect to i) number of trials,
171 ii) mean size of estimation error, and iii) range of actual base rates (BR). Furthermore, the
172 assignment of stimuli to the two conditions (GOOD and BAD), the different estimation error
173 sizes, and the order of trials were randomized anew for each subject. This was accomplished
174 by manipulating the BR, unbeknownst to subjects. To generate a desirable BR, a number
175 between 1 and 25 was subtracted from the estimated BR (eBR), and to generate an
176 undesirable BR, a number between 1 and 25 was added to the eBR. In addition, BRs were
177 capped between 1 and 90% because base rates exceeding this range are likely to appear
178 implausible.

179 However, this manipulation of BR was sometimes constrained by subjects' responses.
180 For instance, when the eBR was close to or greater than 90%, trials that were scheduled to
181 generate bad news could not be realized (e.g., when eBR was 90%, no greater BR could be
182 generated because BRs were capped between 1% and 90%). Instead, a number lower than
183 90% was presented (a random number between 85% and 90%), generating good news. This
184 reversal of the scheduled trial valence (M = 1.30, SD = 2.20) can be made responsible for the

condition-wise differences in number of trials (the reversal occurred only in BAD trials, thereby decreasing the number of realized BAD trials and increasing the number of realized GOOD trials), and eBR and E1 (only BAD trials with high eBRs were possible candidates for such a reversal, and eBR and E1 were highly correlated, as one would expect, see **Figure 2E-F**). Supporting this assumption, number of trials and eBR differed between the conditions only in subjects with reversals (and/or EE = 0; $t(11) = 2.68$, $p = .022$, eBR, $t(11) = 6.86$, $p < .001$), but not in subjects without such irregularities ($t(11) = 0.32$, $p = .755$, eBR, $t(11) = 2.06$, $p = .064$). Notably, we nevertheless achieved satisfactory balanced distributions of number of trials, eBR, E1, and EE between conditions (e.g., on average 78.17 of 80 trials could be realized, and the mean difference between GOOD and BAD was 1.75 trials; see Table 1). Additional details on the experimental design and the BR manipulation algorithm have been described elsewhere (Kuzmanovic and Rigoux, 2017).

Before the experiment, all participants received written instructions, and completed six practice trials with stimulus events not used in the experiment. In a final debriefing after the experiment, a funneled procedure was used to ensure that subjects did not suspect the manipulation of the base rates, or the purpose of the study. All procedures were in accordance with the World Medical Association Declaration of Helsinki and were approved by the local ethics committee of the Medical Faculty of the University of Cologne, Germany (15-255).

Acquisition parameters. The MRI data were acquired by using a Magnetom Trio Prisma^{fit} 3T whole body scanner and a 64-channel head coil (Siemens AG, Medical Solutions, Erlangen, Germany). During the update experiment, fMRI data were acquired in one session with a slice accelerated multiband echo planar imaging sequence (Xu et al., 2013) covering the whole brain (TR = 1050 ms, TE = 37.40 ms, field of view = 212 × 212 × 144 mm^3, voxel size = 2 × 2 × 2 mm^3, 72 oblique axial slices, multiband acceleration factor 6). In addition, we acquired two images with reversed phase encoding directions (anterior-posterior or posterior-

anterior) for the purpose of estimating and correcting the susceptibility-induced distortion using topup (TR 8240 ms, TE 69 ms, field of view 212 × 212 × 144 mm^3, voxel size 2 × 2 × 2 mm^3, 72 oblique axial slices). High-resolution T1-weighted images were obtained from the institute's subject database (MDEFT, TR 1930 ms, TE 5.80 ms, field of view 256 × 256 × 160 mm^3, voxel size 1 × 1 × 1.25 mm^3, 128 sagittal slices, or MPRAGE, TR 2300 ms, TE 2.32 ms, field of view 256 × 256 × 192 mm^3, voxel size 0.9 × 0.9 × 0.9 mm^3, 213 sagittal slices).

Statistical Analyses

Analysis of Task Performance

Prior to analyses, the following trials were excluded: trials with missing responses (M = 0.83, SD = 1.01), trials with EE = 0 (e.g., when eBR was 1% in a GOOD trial, BR was also 1%; M = 0.88, SD = 1.26), and outliers (trials in which the update exceeded 4 SD of the subjects' mean; M = 0.13, SD = 0.34). For each subject, trials were divided into two conditions: *good news* (GOOD; BR < eBR) and *bad news* (BAD; BR > eBR). Optimism bias was assessed by comparing updates in GOOD-trials with those in BAD-trials (mean $update_{GOOD}$ – mean $update_{BAD}$). Note that on average participants were expected to decrease their risk estimates after good news and to increase their risk estimates after bad news. Thus, for both $update_{GOOD}$ and $update_{BAD}$, positive values indicate an update towards the new information (see **Table 1** for statistics of task variables; also see **Figure 2A**). Furthermore, for each participant, we conducted a linear regression to predict his or her updates on each trial using valence of news (GOOD vs. BAD), while including eBR, E1, and EE as covariates (all measures z-scored within subject). For repeated measures, the standard deviation of the paired differences was used as a standardizer for Cohen's *d* (Cumming, 2014).

In addition, we performed computational modeling of belief updating. The model-based approach allows to formally control for fluctuations in trial-wise eBR, E1, and EE across conditions, and to simulate unbiased updating based on observed trial-wise EE and personal relevance (PR). Building on previous work (Kuzmanovic and Rigoux, 2017), the model of belief updating was formalized as follows:

$$Update = LR * EE * PR$$

$$\text{With } LR_{GOOD} = Alpha + Asymmetry$$

$$LR_{BAD} = Alpha - Asymmetry$$

This model relies on the generic form of reinforcement learning, in which update is proportional to the *estimation error* EE (equivalent to prediction error). In addition, EE is weighted by the *learning rate* (LR), which indicates the general tendency of each subject to update their beliefs in response to the EE. In order to test for the optimism bias (asymmetric learning), LR was estimated separately for good and bad news (see also Palminteri et al., 2016; Lefebvre et al., 2017) and therefore has two components. The general component, *Alpha* (α), indicates the tendency to learn from errors, independent of the valence of news. *Alpha* equal 1 indicates that update is exactly equal to EE, while α smaller than 1 indicates updates smaller than EE. *Asymmetry* (A) equal zero indicates equal learning rates for GOOD and BAD, while A different from zero indicates that the resulting learning rates systematically differ for GOOD and BAD (e.g., *Asymmetry* > zero indicates lower learning rates and hence smaller updates for BAD than for GOOD).

EE is also weighted by the *personal relevance* (PR; corresponds to 'relative personal knowledge' in Kuzmanovic and Rigoux, 2017). PR indicates the difference between eBR and E1 relative to the maximal possible difference in each trial (see **Table 1** for the exact equation). Recently, we have demonstrated that the computational model of belief updating

that weighted EE with PR was superior to the model without any consideration of PR (Kuzmanovic and Rigoux, 2017). This shows that the more people felt detached from the reference population, the more irrelevant base rates became for their updates of risk estimates (e.g., if I do not have a car, I will not consider the base rate of car theft). PR ranged from 0 to 1, with PR = 1 when a subject perceives her risk to be equal to those of the average person (eBR = E1, see **Figure 1** for an example), and PR = 0 when the perceived difference (eBR vs. E1) is maximal. Thus, EE weighted by PR indicates a subjective error (SE), where the impact of the EE on update is also determined by the personal relevance of the new information.

Using the VBA toolbox (Daunizeau et al., 2014), we implemented competing models and tested which of these best accounted for the observed update behavior. In order to test whether *Alpha* was different from 1 and whether *Asymmetry* was different from 0, we generated all possible variations of the update equation by switching the parameters *Alpha* and *Asymmetry* on (by letting the parameter free), or off (by fixing the parameter's prior variance to zero). Thus, 4 models (αA, α, A, Ø; α and A indicate that the respective parameter was let free) were estimated for each subject. Note that by setting *Asymmetry* to 0 (i.e., models α and Ø) we specified the null hypothesis that learning is unbiased. In the alternative hypothesis (i.e., models αA and A), *Asymmetry* was estimated for each participant. Model estimations yielded a posterior distribution across the parameters, and an approximation to the evidence of the model. The approximated model evidence reflects the goodness of fit, penalized for the complexity. We used the Free-energy approximation that has been shown to be superior to other approximations like AIC or BIC (Penny, 2012). Model evidences of all subjects and all tested models were then entered in a random-effect Bayesian model comparison. For each model, this procedure estimates a) the probability of each subject to be best described by the respective model (model attributions), b) the frequency in the population (estimated model frequency, *Ef*), and c) the protected exceedance probability (*pxp*), which is

the probability that the model predominates in the population, above and beyond chance (see Rigoux et al., 2014 for more details).

fMRI Analyses

Prior to analysis, the first ten volumes were discarded to allow for magnetic saturation. First, functional images were corrected for motion and distortion using the FSL (version 5.0.9) tools MCFLIRT and topup (Andersson et al., 2003; Smith et al., 2004). All further analysis steps including DCM were conducted using SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK) implemented in MATLAB R2014b (The Mathworks Inc., Massachusetts, USA). The T1 image was normalized to the Montreal Neurological Institute (MNI) reference space using the unified segmentation approach, and the ensuing deformation parameters were applied to (previously coregistered) functional images. Finally, functional images were smoothed using an 8 full-width-half-maximum Gaussian kernel.

Statistical analyses were conducted in the framework of a general linear model (GLM). At the single-subject level, conditions were modeled using a boxcar reference vector convolved with the canonical hemodynamic response function and its time derivative. Following events were modeled on separate regressors: eBR, E1, BR, E2, responses and rest. The duration of eBR, E1, BR, and E2 was always set to 2 s, as for events with responses (all except of BR) the response display was activated only after 2 s. Responses for all events were modeled on one regressor (duration from the onset of the response event to the confirmation button press, which was also the beginning of the next inter-stimulus-interval). The instruction to switch hands after the first half of trials, and the excluded trials (missing responses, EE = 0, and outliers; see *Analysis of Task Performance*), if present, were modeled on the 'rest' regressor. Motion parameters and a matrix with motion-outlier volumes (identified using the tool fsl_motion_outlier at a threshold of 4 *SD* of intensity differences

313 between subsequent volumes (Power et al., 2012)) were included as multiple regressors of no
314 interest. Low-frequency signal drifts were filtered using a cutoff of 128 s. At the group-level,
315 flexible factorial design and a significance threshold of $p < .05$, FWE-corrected at the peak
316 level, with an extent threshold of 20 voxels were used. For the covariate analyses, we applied
317 the same statistical threshold, but a lower extent threshold of 10 voxels.
318
319 *Error Tracking.* We identified brain regions that encoded the errors experienced
320 during the BR event. At that time, subjects were confronted with a different actual base rate
321 than the one they have estimated (i.e., the difference between eBR and BR). In order to obtain
322 the effects separately for GOOD and BAD, we split the BR trials into BR_{GOOD} and BR_{BAD},
323 and tested for parametric modulation (PM) by subjective error (SE = EE * PR, see *Analysis of*
324 *Task Performance*). We focused on this subjective error processing because it was more
325 relevant for the subsequent belief updating than the general error (i.e., EE; Kuzmanovic and
326 Rigoux, 2017). The resulting 9 regressors (eBR, E1, BR_{GOOD}, PM_error_{GOOD}, BR_{BAD},
327 PM_error_{BAD}, E2, responses, and rest) were only weakly correlated ($\bar{r}$s, averaged across
328 subjects, between -.29 and .14), indicating efficient parameter estimation. At the single-
329 subject level, two contrast images were computed relative to the implicit baseline
330 (PM_error_{GOOD} and PM_error_{BAD}) and entered into group-level analysis. At the group level,
331 we identified those regions that exhibited increasing or decreasing activation with increasing
332 subjective error in both GOOD and BAD trials (global conjunction). Furthermore, we
333 explored differences between PM_error_{GOOD} and PM_error_{BAD} and reported global
334 conjunction results for significant results to clarify whether the difference related to different
335 magnitudes of the same modulation effect (e.g., the positive correlation between BOLD and
336 error was stronger in BAD than in GOOD), or to modulation effects of opposite direction
337 (e.g., the correlation between BOLD and error was positive in BAD, but negative in GOOD).
338 In order to be able to illustrate group effects of different sizes of error on the BOLD signal,

we also computed a GLM that models three sizes of error (small, mid and large) on 3 separate regressors, separately for GOOD and BAD (**Figure 3A**, line chart). 'Small', 'mid', and 'large' categories were generated by dividing the sorted array of values into three subarrays that i) do not share same values, and ii) are maximally similar with respect to the number of elements (this procedure was the same for errors and update', see below). Finally, we tested whether the extent of error tracking correlated with the learning rate component *Alpha* across subjects by conducting a covariate analysis with one contrast per subject (average effect of PM_error_{GOOD} and PM_error_{BAD}).

The Valence of Updating. In order to identify brain regions that encoded the valence of updating, we focused on the E2 event because at that time subjects were deciding upon updating their initial belief. E2 trials were split into $E2_{GOOD}$ and $E2_{BAD}$ trials so that effects can be examined separately for GOOD and BAD. According to the valence of updating schematically shown in **Figure 3B** (gray box), we tested for the positive correlation between the BOLD-signal and update in GOOD trials, and for the negative correlation in BAD trials. To identify these opposed effects, we applied parametric modulation of $E2_{GOOD}$ and $E2_{BAD}$, respectively, by update size. The advantage of the PM procedure is that it allows to adjust the effect of update for EE, PR, and other potential confounds (e.g., by including three orthogonalized parameters in the following order: PR, EE, update; Mumford et al., 2015). The resulting 13 regressors (eBR, E1, BR, $E2_{GOOD}$, PM_PR_{GOOD}, PM_EE_{GOOD}, PM_update_{GOOD}, $E2_{BAD}$, PM_PR_{BAD}, PM_EE_{BAD}, PM_update_{BAD}, responses, and rest) were only weakly correlated ($\bar{r}$s, averaged across subjects, between -.29 and .03), indicating efficient parameter estimation. The only exception was the negative correlation between BR and response ($\bar{r}$ = -.49), which occurred because BR was the only event in the trial that was never associated with a subsequent motor response. Six contrast images were computed relative to the implicit baseline (PM_PR, PM_EE, and PM_update, separately for GOOD and BAD), and entered

into group-level analysis. At the group level, we identified those regions that exhibited both i) increasing activation with increasing updates in GOOD trials, as well as ii) increasing activation with decreasing updates in BAD trials, specified by the difference contrast ($PM\_update_{GOOD} > PM\_update_{BAD}$). In addition, we reported global conjunction results to clarify whether the difference related to different magnitudes of the same modulation effect (e.g., the positive correlation between BOLD and update was stronger in GOOD than in BAD), or to modulation effects of opposite direction (e.g., the correlation between BOLD and update was positive in GOOD, but negative in BAD). Moreover, we tested whether the magnitude of the favorable updating effect correlated with the optimism bias across subjects by conducting a covariate analysis with one contrast per subject ($PM\_update_{GOOD} > PM\_update_{BAD}$).

Finally, we conducted two additional GLMs with categorical designs that split all trials into three sizes of update (small, mid, and large), separately for GOOD and BAD. To approximate the adjustment for EE within the modulation by update, we subtracted EE from update at each trial (update' = update – EE). That way, we controlled for the general effect that updates tend to be larger after larger EE, which may confound with the valence effect. Note that dividing update by EE would not be optimal because all trials with an update equal zero (M = 30.09%, SD = 16.13) would have yielded zero as well, irrespective of EE. This would not be appropriate because meaningful differences between zero updates in response to EEs of different sizes (e.g., EE = 2 and EE = 20) would have been concealed. For each subject, the numbers of trials across the three categories of updates were kept as similar as possible (numbers of trials did not differ; GOOD: $M = 13.32$, $SD = 0.85$, $F(2,69) = 0.58$, $p = .560$; BAD: $M = 12.74$, $SD = 1.02$, $F(2,69) = 0.09$, $p = .913$).

First, we used a GLM that modeled the three categories of update sizes separately for GOOD and BAD (six regressors) in order to be able to illustrate group effects of different sizes of updates on the BOLD signal (**Figure 3B**, line chart). Second, we used another

categorical GLM as a basis for the DCM analysis because the categorical levels can be more easily interpreted as inducing contextual modulatory effects in DCM than parametric variables (Stephan et al., 2010). According to the valence of updating schematically shown in **Figure 3B** (gray box), this GLM collapsed the different update sizes into three valence categories corresponding to *unfavorable* ($small_{GOOD}$ and $large_{BAD}$ updates, U), *mid* (mid updates, M) and *favorable* ($large_{GOOD}$ and $small_{BAD}$ updates, F) updating (**Figure 3C**). Three contrast images were computed relative to the implicit baseline ($E2_{unfavorable}$, $E2_{mid}$ and $E2_{favorable}$), and entered into group-level analysis. At the group level, we tested for brain regions that exhibited greater activation for favorable updates than for unfavorable updates ($E2_{favorable} > E2_{unfavorable}$). In addition, we identified those regions that were activated during updating independent of the valence (i.e., conjunction of all three levels of E2).

DCM Analyses

DCM represents a hypothesis-led approach to understand neural circuits underlying observed brain responses (Friston, 2011). We used DCM to estimate and infer causal interactions among brain regions involved in belief updating (i.e., during the E2 event). To this end, competing models with different intrinsic coupling between regions, and different task-dependent modulations of these couplings, were specified. Each model corresponded to a specific hypothesis about how observed data were caused, and Bayesian model selection was used to quantify the evidence for one model over another (Friston, 2011). Model inversion provided estimates of the model evidence and the corresponding effective connectivity. We tested whether the context of favorable and unfavorable updating modulated the coupling between distributed brain responses, and whether value-coding regions exerted influence on other regions associated with cognitive processing.

First, we selected the nodes for the DCM based on the group results revealed by the simplified categorical GLM with three categories of valence of updates (unfavorable, mid and favorable). The time series were extracted by computing the principal eigenvariate from 4 mm diameter spheres (33 voxels) centered on the peak coordinates and adjusted for the effect of interest (F-contrast across the three categories of updates and the respective time derivatives).

Second, we specified competing models varying in their endogenous coupling and valence-dependent modulatory effects and inverted each model for every subject. Given that every brain region is connected reciprocally (Friston, 2011), the coupling in all models was cyclic, i.e., all forward connections were accompanied by respective backward connections. We used a random-effect Bayesian model comparison to infer the optimal model structure by selecting the model with the best balance between accuracy and complexity.

Third, following the model selection, we performed a random-effect analysis of parameter estimates derived from the selected model using one-sample t-tests (Stephan et al., 2010). Additionally, we tested for correlations between parameter estimates and the optimism bias. For the sake of completeness, we report correlations between the two bias measures (optimism bias and Asymmetry) and all model parameters in **Table 4**. Bonferroni correction was used to control for multiple comparisons: significance thresholds were adjusted for 7 tests for the matrix A parameters ($p < .007$), and 2 tests for the expected correlations ($p < .025$).

Results

Task Performance

In the belief update task, subjects were asked to reconsider their risk estimates after being confronted with either good news (base rates of the risks were *lower* than initially expected), or bad news (base rates were *higher* than initially expected; **Figure 1**). In order to assess valence-biased belief updating, we first tested whether subjects were more likely to take into account good news (GOOD) rather than bad news (BAD). Indeed, belief updates (the difference between the self-related risk estimates before and after being presented with the actual base rate) following good news were significantly larger than the updates after bad news, $t(23) = 2.12$, $p = .045$, paired t-test, $d = 0.43$ (**Figure 2A**, see **Table 1** for the summary of all task variables). Furthermore, linear regression analyses revealed that updates were larger in GOOD than in BAD trials even after controlling for trial-wise estimation error (EE), $t(23) = 3.03$, $p = .006$, $d = 0.62$, or for estimated base rate (eBR), first self-related risk estimate (E1) and EE, $t(23) = 2.55$, $p = .018$, $d = 0.52$, one-sample t-tests. Thus, these results indicate that belief updates were optimistically biased.

In order to implement an even more precise control for potential confounds, and to further inform the fMRI analyses, we applied computational modeling. We tested whether the learning from actual base rates was asymmetric (different for GOOD and BAD, indicated by the parameter *Asymmetry*) while taking into account the EE, the personal relevance of the new information (PR), and the general tendency to learn from new information (learning rate component *Alpha*). Here, the EE was an important confound because larger errors generally tend to trigger larger updates. Also, when the new information is not regarded as personally relevant, updating of related beliefs tends to be reduced.

Bayesian model comparison of four competing models (αA, α, A, Ø) provided additional support for the optimism bias. It revealed that the 'αA'-model, that estimated both *Alpha* and its *Asymmetry* separately for each subject, predicted subjects' behavior significantly better than all other model versions (α, *Alpha* fitted, *Asymmetry* fixed to 0; A, *Alpha* fixed to 1, *Asymmetry* fitted; or Ø, *Alpha* fixed to 1 and *Asymmetry* fixed to 0), Ef = .87, *pxp* = .994 (**Figure 2B**). *Asymmetry* was significantly larger than zero, M = 0.05, SD = 0.07, $t(23) = 3.59$, $p = .002$, one-sample t-test, $d = 0.73$, showing that participants' learning rates were higher in response to good news than to bad news (LR_{GOOD}: M = 0.79, SD = 0.19, LR_{BAD}: M = 0.70, SD = 0.18; **Figure 2C**). Furthermore, *Alpha* was significantly smaller than 1, M = 0.74, SD = 0.18, $t(23) = -7.18$, $p < .001$, one-sample t-test, $d = 4.26$, showing that updates were on average smaller than the estimation errors. Finally, the optimism bias (derived from the observed task performance, mean $update_{GOOD}$ – mean $update_{BAD}$) and the *Asymmetry* parameter (derived by the winning model 'αA') were significantly correlated, r = .79, $p < .001$ (**Figure 2D**). While expected, the close relationship between these two bias measures also confirmed that the potential confounding variables (EE, PR) had no systematic influence in our task. Thus, we can rule out that "seemingly optimistic updating" was induced by a differential consideration of estimation errors due to varying personal relevance (Shah et al., 2016, p. 92). Quite the contrary, the optimism bias was even stronger after taking EE and PR into account. Thus, it is likely that earlier studies demonstrating the optimism bias, but lacking the enhanced experimental or formal computational control (Sharot et al., 2011; Sharot et al., 2012; Garrett et al., 2014; Korn et al., 2014; Kuzmanovic et al., 2015, 2016a; Kuzmanovic et al., 2016b), are also not affected by these potential confounds. Furthermore, correlations between the different task variables, computed separately for trials with good news and bad news and then averaged across subjects (**Figure 2E-F**), show that EE and updates correlated only very weakly with eBR, BR, E1, and E2 ($\bar{r}$s ranging from -.15 to .30). This is particularly important as it demonstrates that we succeeded in manipulating the

485 desirability of EE independently of prior beliefs (i.e., the size of risk estimates eBR and E1).
486 Furthermore, it shows that the valence of updates was independent of the size of the estimated
487 risks (eBR, E1) or the presented base rates (BR). Together, these findings provide a strong
488 support for the notion that the difference in updating indeed reflected a valence-dependent
489 consideration of the new information.

490 Moreover, we assessed the updates that were simulated by the winning model 'αA'
491 assuming asymmetric learning rates, and by the unbiased model 'α', given the trial-by-trial
492 PR and EE. The updates simulated by the model 'αA' corresponded well to the actually
493 observed updates (M_{GOOD} = 7.59, SD = 2.43; M_{BAD} = 6.49, SD = 2.09, see white dots in
494 **Figure 2A**) and were larger in GOOD than in BAD trials, $t(23) = 3.85$, $p = .001$, paired t-test,
495 $d = 0.79$. In contrast, the updates simulated by the unbiased model 'α' did not differ across
496 GOOD and BAD trial (M_{GOOD} = 7.06, SD = 2.21; M_{BAD} = 7.02, SD = 2.20, $t(23) = 0.22$, $p =$
497 .830, paired t-test, $d = 0.04$; see gray dots in **Figure 2A**). This comparison proves that
498 subjects' asymmetric updating represents a true bias attributable to the different valence of the
499 new information (good and bad news) and cannot be explained by any variations of other
500 trial-by-trial variables (i.e., PR or EE; see Shah et al., 2016; Kuzmanovic and Rigoux, 2017).

501 Furthermore, we compared floor and ceiling effects across GOOD and BAD and
502 showed that controlling for these effects even enhanced the optimism bias effect. Floor and
503 ceiling effects could occur if the size of the possible update was limited by the response scale
504 (probabilities from 1% to 99%). For example, in a GOOD trial, given an EE = 5 (e.g., eBR =
505 10%, BR = 5%) and an E1 = 3%, a subject would have only a limited space on the response
506 scale to make an update toward a lower risk estimate (from E1 = 3% to the end of the
507 response scale of 1%). Critically, this possible update should be at least as large as the size of
508 the estimation error in order to enable unconstrained updating. This is rather conservative
509 because the general learning rate component *Alpha* was significantly smaller than 1 and
510 because EE were also weighted by personal relevance that ranged between 0 and 1. In order to

test for floor and ceiling effects, we computed the size of possible update relative to EE for each trial (updspace-EE_{GOOD} = (E1-1) - EE; updspace-EE_{BAD} = (99-E1) - EE). Updspace-EE was lower in GOOD than in BAD, t(23) = -4.50, p < .001 (M_{GOOD} = 28.24, SD = 11.03; M_{BAD} = 47.31, SD = 10.51). Furthermore, the number of constrained update spaces (updspace-EE < 0) was higher in GOOD than in BAD, t(23) = 5.08, p < .001 (M_{GOOD} = 6.17, SD = 5.45; M_{BAD} = 0.58, SD = 0.77). Repeating the analyses after excluding the trials with a constrained update space revealed an even stronger optimism bias effect, $t(23) = 3.54$, $p = .002$, paired t-test, d = 0.72, M_{GOOD} = 8.20, SD = 2.64; M_{BAD} = 6.80, SD = 2.23, also when after controlling for trial-wise EE, $t(23) = 4.67$, $p < .001$, $d = 0.95$, or for eBR, E1 and EE, $t(23) = 4.10$, $p < .001$, d = 0.84, one-sample t-tests. Computational modeling analyses were not affected by the exclusion of trials with constrained updating due to formal consideration of the personal relevance: *Asymmetry* derived from the 'αA'-model (Ef = .86, pxp = .991) was significantly larger than zero, $t(35) = 3.50$, $p = .002$, one-sample t-test, $d = 0.72$. Taken together, these tests show that the optimism bias effect was underestimated because of greater floor effects in GOOD trials.

We also examined the behavioral results after adding the 12 subjects with excessive motion to the sample of 24 subjects. These analyses yielded the same results as those with n=24. Belief updates following good news were significantly larger than the updates after bad news, $t(35) = 3.18$, $p = .003$, paired t-test, $d = 0.53$; *Asymmetry* was significantly larger than zero, $t(35) = 4.35$, $p < .001$, one-sample t-test, $d = 0.73$. Finally, the post-experimental debriefing revealed that none of the included subjects suspected that the purpose of the task was to assess difference in belief updating depending on the valence of the new information. At the end of the debriefing, we carefully explained the purpose of the study as well as the manipulation of the base rates in a standardized written form. Following this information, only one subject reported that he was aware of the good news-bad news effect during the own task performance. Furthermore, two subjects reported that they had no concerns with respect to the presented base rates during the task, the majority (17) reported that they were surprised

by some of the presented base rates but did not doubt their validity, five subjects doubted that single surprising base rates were really valid, and none of the included subjects reported having realized that the base rates were manipulated.

fMRI Results

Error Tracking During BR. Updating beliefs about self-related risks was triggered by erroneous expectations regarding the respective base rates. In order to investigate this crucial process, we identified brain regions that tracked the errors experienced upon the presentation of the actual base rates. The parametric modulation (PM) analysis revealed that error tracking recruited the anterior cingulate cortex (ACC), the inferior frontal gyrus (IFG), the anterior insula, the middle orbital gyrus and the dorsolateral prefrontal cortex (dlPFC; **Table 2** contrast 1b, and **Figure 3A)**. In these regions, the activity increased with decreasing error size (negative correlation between BOLD and error) for both conditions GOOD and BAD (for an example, see the line chart in **Figure 3A** for the average activity in dlPFC across three sizes of error). This negative correlation seems unexpected because brain regions such as the ACC and the anterior insula have been associated with error processing, novelty, and task difficulty (Wessel et al., 2012; Klein et al., 2013; Shenhav et al., 2014; Kolling et al., 2016; Bastin et al., 2017; Fouragnan et al., 2017), and thus were expected to increase activity with increasing error size. However, a seminal study on belief updating has also demonstrated negative correlation between estimation error size and activity in the IFG, which was moreover predictive of trait optimism (Sharot et al., 2011). Thus, it may be necessary to reconsider the meaning of different outcomes in the specific context of the present experiment, because contexts determine the reference point for values of options (Palminteri et al., 2015). Subjects' task and their 'default option' was to revisit their prior beliefs due to new challenging information. Thus, in the majority of trials, subjects indeed were confronted with

base rates that markedly differed from what they expected and they updated their belief. Relative to this, encountering trials with a small error increases the difficulty of the decision whether to update beliefs ("Is the actual base rate different enough than expected, and is this difference relevant enough to drive an update of my own risk?"). Given the high accuracy of subjects' base rate estimations in such trials, the alternative course of action to refrain from updating becomes increasingly valuable. We therefore speculate that increased activity in this network relates to enhanced initial comparison process that informs subsequent decisions about updating while maintaining behavioral flexibility (Kolling et al., 2016).

Furthermore, of all these error-tracking regions, only the activity in dlPFC correlated with the learning rate component *Alpha* (covariate analysis masked with the conjunction contrast PM_error$_{GOOD}$ and PM_error$_{BAD}$, negative correlation; **Table 2** contrast 2). Betas indicating the strength of the linear relationship between error and BOLD were extracted for each subject (PM analysis, at the peak [40 38 30], averaged across PM_error$_{GOOD}$ and PM_error$_{BAD}$), and plotted against *Alpha* for illustrative purposes (scatter plot in **Figure 3A**). Even when conducting the covariate analysis for the whole brain, an overlapping dlPFC cluster had the strongest correlation with *Alpha* (peak at [38 30 38], T = 5.66, 257 voxels), albeit at a more liberal significance threshold ($p < .05$, FWE-corrected at the cluster level).

The Valence of Updating. The main aim of the fMRI analysis was to identify brain regions that encoded the valence of updating. The valence of updating was defined based on how much the second estimation resulted in either favorable or unfavorable risk estimates, relative to the first estimation (see **Figure 3B**, gray box, for an illustration). Note that in GOOD trials, initial risk estimates were expected to *decrease* toward the actual base rate that was lower than expected. Conversely, in BAD trials, risk estimates were expected to *increase* toward the actual base rate that was higher than expected. In consequence, for GOOD trials, we assume that *large* updates would be experienced as *favorable*, because they result in lower

589 final risk estimates. In contrast, for BAD trials, we assume that *small* (or zero) updates would
590 be experienced as *favorable*, because they prevent an increase of final risk estimates. Hence,
591 we expected a *positive correlation* between the BOLD-signal and update in GOOD trials, and
592 *a negative correlation* in BAD trials.

593 The parametric modulation analysis revealed that activity in the vmPFC had exactly
594 this pattern (**Figure 3B** and **Table 3**, contrast 1a), indicating that this region tracked favorable
595 updating. The correlation between the BOLD-signal in the vmPFC and update was greater in
596 trials with good news than in trials with bad news (i.e., $PM\_update_{GOOD} > PM\_update_{BAD}$).
597 The conjunction contrast (i.e., PM_update_{GOOD}, positive correlation & PM_update_{BAD},
598 negative correlation) confirmed that this effect implied contrary modulation effects for GOOD
599 and BAD (positive correlation in GOOD, and negative correlation in BAD; see the line chart
600 in **Figure 3B**). Moreover, the vmPFC was also the only area in the whole brain, in which the
601 magnitude of this valence-tracking effect correlated with the optimism bias (**Table 3**,
602 contrasts 2a-b). This relationship to the task performance was illustrated by extracting betas
603 indicating the strength of the differential linear relationship between update and BOLD for
604 each subject (PM analysis, at the peak -6 50 -18, $PM\_Error_{GOOD} > PM\_Error_{BAD}$), and
605 plotting them against optimism bias (**Figure 3B**, scatter plot).

606 Importantly, by including multiple orthogonalized parameters, we assessed variance
607 that was uniquely explained by different update sizes, above and beyond the effects of other
608 relevant computational components of belief updating such as EE and PR (Mumford et al.,
609 2015). Moreover, the valence-tracking effect in the vmPFC was significant even when we
610 controlled for task variables other than EE and PR. Repeating the parametric modulation
611 analysis while including BR as an additional regressor (four orthogonalized regressors: BR,
612 PR, EE, update, separately for GOOD and BAD) yielded the involvement of the same vmPFC
613 clusters for the three contrasts indicating the valence effect (see Table 3, 1a including the
614 conjunction and 2b; significance threshold as in the main analysis). In addition, including BR

and E2 (five orthogonalized regressors: E2, BR, PR, EE, update, separately for GOOD and BAD) also confirmed the valence-tracking effect in the vmPFC with respect to all three contrasts (albeit the contrast 2b at a less stringent significance threshold of $p < .001$, uncorrected, cluster size 216). Thus, in contrast to previous studies investigating the rewarding effect of favorable new information per se (in the context of updating self-evaluations; Korn et al., 2012), the valence effect related to the *relative improvement or worsening of initial beliefs*, but not to the valence of final beliefs (E2) or the new information (BR), or to other variables (PR, or EE).

Once we have demonstrated the effect of favorable updating while adjusting for BR, E2, EE and PR using parametric modulation, we repeated the analysis with a simplified categorical model. Discrete levels of valence (e.g., unfavorable or favorable) can then be used in the following DCM analysis to specify contextual effects that modulate the intrinsic coupling within the update circuit. According to the principle introduced above (and in the gray box in **Figure 3B**), we specified three valence levels (**Figure 3C**): *unfavorable* updating (U, small$_{GOOD}$ and large$_{BAD}$ updates), *mid* updating (M, mid updates) and *favorable* updating (F, large$_{GOOD}$ and small$_{BAD}$ updates). As a result, we concatenated the valence of updates across trials with good and bad news. The simplified categorical analysis revealed similar results as the PM analysis, demonstrating greater activity in the vmPFC for favorable than for unfavorable updates (**Figure 3C** and **Table 3**, contrast 3a). Moreover, the dmPFC showed a similar pattern of activity as the vmPFC. In addition, we tested for the conjunction effect across all three valence categories to identify brain regions that were generally activated during updating, independent of valence. General updating revealed widespread activations including occipital, parietal and frontal cortices (**Figure 3C** and **Table 3**, contrast 3c).

The Timing of Updating. In order to explore whether the valence effect of updating did manifest already during the processing of BR, we repeated the parametric modulation analysis

641 (including PR, EE, and update), but defined the event BR instead of E2 as the unmodulated
642 regressor. In this analysis, the contrast PM_update$_{GOOD}$ > PM_update$_{BAD}$ did not yield any
643 significant effect, even when using the vmPFC cluster as an inclusive mask at a less stringent
644 significance level ($p < .001$, uncorrected). This indicates that the encoding of the valence of
645 belief updating by the vmPFC indeed occurred during the period of update consideration and
646 not already during the reception of the new information.
647 This contradicts classical reinforcement tasks where belief updating is expected to
648 occur upon a relevant outcome (e.g., if I choose the green and not the red square and win
649 money, I update the value of choosing the green square immediately). However, there is a
650 substantial qualitative difference between classical reinforcement tasks and the present task.
651 Estimations of base rates of life events in a population recruit declarative memory to retrieve
652 general knowledge, and the feedback about the actual base rates indicates how accurate one
653 was. Furthermore, estimating one's own risks of experiencing adverse events in the future
654 represents a more complex cognitive process (including autobiographic and declarative
655 memory) than deciding whether to choose a green or a red square in a gambling game. Thus,
656 it is plausible that subjects focused on their degree of accuracy at the time point of feedback
657 and subsequently focused on the meaning of this new information for their own risk estimate.
658 Several arguments additionally support this notion. First, we explicitly instructed the
659 subjects to reconsider their risk estimates during the "update phase" (first two seconds of E2,
660 before the response buttons were activated, see Figure 1). Second, there was no need to
661 memorize estimated and actual base rates (or first self-risk estimates) until the update phase
662 because all preceding values were visible on the screen at all times (see Figure 1). Third,
663 during the debriefing, a majority of subjects spontaneously reported that they were pleased to
664 see that they were often quite accurate in estimating the base rates. And forth, inspecting the
665 encoding of estimation errors upon presenting the actual base rates (i.e., parametric
666 modulation of actual base rate presentation by orthogonalized PR and EE) revealed that the

activity in the bilateral ventral striatum was higher the smaller the estimation errors were (the more accurate the subjects were), irrespective of their personal relevance or desirability (i.e., for both good news and bad news; left striatum [-12 14 -6], 68 voxel; right striatum [14 12 -6], 73 voxel; $p < .05$, FWE-corr. at the peak level for the whole brain). The ventral striatum plays a central role in encoding positive prediction errors (Chase et al., 2015). In the context of a task where subjects estimated base rates and were confronted with actual base rates that differed from their own estimates to a varying extent, greater accuracy corresponded to positive prediction error. Together with the debriefing self-reports, this finding supports the assumption that subjects focused on the degree of their accuracy during the presentation of base rates and thus were likely to reconsider their own risks subsequently, at a segregated time point.

DCM results

After identifying the vmPFC as the valuation area in the context of belief updating, we applied DCM to test competing hypotheses about its causal role within the update circuit. First, we selected three nodes for the DCM based on the update-related group results revealed by the simplified categorical analysis (**Figure 3C** and **Table 3**, contrasts 3a-c). The first node was the dlPFC. This region was involved both in general updating (conjunction across all update categories, **Figure 3C**), as well as in tracking errors in relevant prior beliefs (regarding BR) that was predictive of individual learning rates (**Figure 3A**). Thus, within the network activated by general updating, we chose the peak nearest to the learning rate-associated error tracking effect (MNI peak coordinate [44 42 26]). In order to further ensure that the chosen dlPFC peak was indeed specifically recruited by updating, we contrasted the three E2 categories with E1 (second versus first self-risk estimation). In a separate group-level analysis with four contrast images ($E2_{unfavorable}$, $E2_{mid}$ and $E2_{favorable}$, E1) we identified those regions

693 that were more activated during updating beliefs about risks than during forming initial beliefs
694 about risks (i.e., $E2_{unfavorable}$, $E2_{mid}$ and $E2_{favorable}$ > E1, contrast [1 1 1 -3]). Note that E1 and
695 E2 were otherwise comparable with respect to visual and motor requirements (see Figure 1).
696 This additional analysis confirmed that the dlPFC was significantly activated during belief
697 updating relative to initial belief formation ($p < .05$, FWE-corr. at the peak level for the whole
698 brain). Importantly, the dlPFC cluster overlapped with both the general updating conjunction
699 effect and the error tracking effect. Given that the dlPFC is a crucial part of the working
700 memory system for transient storage and manipulation of information (Eriksson et al., 2015),
701 this region represents a key candidate for maintaining integrative information processing
702 generally necessary for belief updating. We therefore refer to the dlPFC as the valence-
703 independent "update processing" node. The second node was the vmPFC (MNI peak
704 coordinate [-2 46 -22]). Its activity was greater in response to favorable than unfavorable
705 updates, and this valence-coding effect predicted the individual magnitudes of the optimism
706 bias (**Figure 3B-C**), thus forming a "valuation" node. Finally, we defined the dmPFC as our
707 third node (MNI peak coordinate [-16 44 40]). This region demonstrated a similar activity
708 pattern as the vmPFC in the categorical GLM (**Figure 3C**), but, in contrast to vmPFC, it has
709 been associated with cognitive processes such as social inferences and perspective taking, and
710 less so with reward processing (Bzdok et al., 2013; de la Vega et al., 2016). Thus, we will
711 tentatively refer to the dmPFC as the "cognitive" node.
712 This simple architecture comprising three nodes allowed us to compare different
713 models that implied either valence-guided or non-valence-guided explanations for the
714 observed brain responses. Valence-guided explanations would be favored if the vmPFC were
715 the source of both i) the valence-dependent filtering of the general update-processing signal
716 and ii) the subsequent influence on other prefrontal regions. Alternatively, non-valence-
717 guided explanations would be supported if the vmPFC would i) receive a signal that is already
718 modulated in a valence-dependent manner and ii) have no driving influence on other

prefrontal regions. Thus, adopting a hypothesis driven approach, we limited our model space to ten DCMs corresponding to these competing theories about the neural processing of belief updating (**Figure 4A**). Although we could in principle construct a higher number of possible models, including more models would mainly obfuscate our analysis as additional alternative models would not be realistic (e.g., disconnected nodes), or be prone to overfitting while unable to provide a conclusive answer to our research question (e.g., valence modulates all connections).

In all ten models, the event corresponding to the second risk estimation (E2, all three categories of updates) was specified as the exogenous input (**Figure 4A**). This input entered the dlPFC (matrix C in DCM) because this region showed increased activity during all categories of belief updating (see the line chart in **Figure 3C**). The models differed in their endogenous coupling (matrix A in DCM) such that *m1* to *m5* assumed a flow of neuronal states from dlPFC via vmPFC to dmPFC, while in models *m6* to *m10* the flow was from dlPFC via dmPFC to vmPFC. Given that we expected that the vmPFC would influence the dmPFC, models *m6* to *m10* represented null-hypotheses assuming the opposite course of influence. Furthermore, we systematically selected each of the possible coupling parameters to be the target of the valence-dependent modulation (matrix B in DCM; unfavorable updating, U or favorable updating, F). Given that we expected that favorable and unfavorable updating would differentially modulate the coupling in the network, *m5* and *m10* represented null-hypotheses assuming no modulation at all. More specifically, we hypothesized that valence encoding would manifest through filtering of the incoming signal by the vmPFC, and that the resulting differential valuation would further influence dmPFC, as formalized in *m1*. Alternatively, the valence-dependent modulation could have affected one of the other couplings (e.g., from dlPFC to dmPFC as formalized in *m6*). In these cases, the filtering of the incoming signal would not be attributed to vmPFC, and/or there would be no primary influence of the vmPFC on dmPFC. Notably, all models except of *m5* and *m10* were equally

complex but differed with respect to the flow of neuronal states and the coupling which was subject to contextual modulation, allowing for evidence-based hypothesis testing.

Bayesian model comparison confirmed that the model *m1* had the greatest evidence, above and beyond chance, *Ef* = .77, *pxp* = .999 (**Figure 4B**). The selected model assumed a cyclic signal flow from the dlPFC via vmPFC to dmPFC and a valence-dependent modulation of the coupling from dlPFC to vmPFC. The low evidence of models *without* modulations (*m5* and *m10*) indicates that the valence-dependent modulation of effective connectivity was indeed necessary to adequately predict subjects' network activity. Furthermore, of the eight models *with* modulations, *m1* still had a profoundly higher evidence than *m6* and other alternative models. This finding supports the hypothesis that the vmPFC filtered the incoming signal in a valence-dependent manner and influenced the dmPFC.

Third, we further inspected and analyzed the parameter estimates derived from the winning model *m1*. We hypothesized that the magnitude of valence-dependent modulation (the difference between F and U, F-U) of the dlPFC-vmPFC coupling would correlate with the optimism bias across subjects, because the stronger this modulation, the greater should be the response of the vmPFC to different valences of updating. Furthermore, we expected that the strength of the connection from vmPFC towards dmPFC would also correlate with the optimism bias, assuming that this coupling represents the influence of valuation on ongoing cognitive processing. Modulation parameters and coupling patterns in the context of favorable and unfavorable updating are reported in **Table 4** and plotted in **Figure 4C**. On average, the coupling from dlPFC to vmPFC decreased in the context of unfavorable updating relative to favorable updating. However, this difference did not reach significance due to the large variance of modulation estimates for U and F. It is of greater importance though that the valence-dependent modulation of the dlPFC-vmPFC coupling (the difference between F and U, F-U) correlated with the size of the optimism bias across subjects (**Figure 4D**). This relationship explains how the data observed in the fMRI analysis were caused. Subjects with a

771 greater optimism bias had a stronger valence-dependent filtering by vmPFC, resulting in an
772 increased BOLD response to favorable than unfavorable updating in vmPFC. Moreover, the
773 individual strength of the coupling from vmPFC to dmPFC also correlated with optimism bias
774 (both correlations corrected for multiple comparisons). Thus, the stronger the optimism bias,
775 the stronger was the influence of valuation on ongoing cognitive processing, mediated by the
776 coupling from vmPFC to dmPFC. In addition, the inspection of all possible correlation
777 coefficients (**Table 4**) revealed that the endogenous self-connection of the dmPFC inversely
778 correlated with the optimism bias (not corrected for multiple comparisons). Thus, the stronger
779 the optimism bias, the weaker was the self-inhibition of the dmPFC. Taken together, these
780 parameter estimates indicate a self-enhancing cyclic flow between vmPFC and dmPFC. In
781 subjects with high optimism bias, the vmPFC filtered the incoming information dependent on
782 valence. This differential signal was then forwarded to the dmPFC and there enhanced by the
783 reduced self-inhibition.

Discussion

The present study provides converging evidence that the value of desirable beliefs can influence ongoing cognitive processing. Participants demonstrated an optimism bias as they were more likely to update beliefs regarding their risks in response to good news than bad news (learning that base rates of the risks were *lower* versus *higher* than expected). This finding was also confirmed by computational modeling that formally controlled for valence-unrelated variables that influence updating (estimation error and personal relevance of the new information; Shah et al., 2016; Kuzmanovic and Rigoux, 2017). Given that we i) ruled out these alternative, valence-independent explanations, ii) manipulated the desirability of the new information independently of prior beliefs, and iii) demonstrated that the optimism bias was unrelated to the size of risk estimates or base rates, we conclude that information integration was indeed biased by the motivation to adopt the most favorable beliefs about one's future.

Furthermore, fMRI results showed that the vmPFC tracked the value of updating. In the context of good news, large updates toward lower risk estimates improve the ultimate risk perception, but after bad news, large updates toward higher risk estimates worsen the ultimate belief. In turn, small updates (small or no change in beliefs) also acquire opposing values in the context of good and bad news, respectively: While small updates after good news are unfavorable as they disregard the opportunity to improve risk estimates, small updates after bad news are favorable because they prevent worsening of risk estimates. The activity pattern in the vmPFC showed exactly this pattern: it increased with *increasing* updates toward lower risks (after good news), and with *decreasing* updates after bad news. Thus, not only improving beliefs, but also avoiding the worsening of beliefs triggered the vmPFC activity. Previous studies on optimism bias that included risk estimates for self and a similar other person already indicated a positive value of avoiding threatening belief updates by

disregarding undesirable new information. Here, particularly the decreased updating in self-related trials with bad news (relative to a comparably high amount of updating in self-related trials with good news and all other-related trials) was driving the optimism bias (Kuzmanovic et al., 2015, 2016a). Furthermore, research on context dependency of option values has shown that both gaining a reward (i.e., improving the current state, e.g., change from 1$ to 2$) and knowingly avoiding punishment (i.e., current state is unchanged, e.g., 1$, but the possible loss, e.g., -1$, is avoided) acquired a positive value and were tracked by the vmPFC (Palminteri et al., 2015). Lesion studies (Camille et al., 2011) and meta-analyses (Yarkoni et al., 2011; Diekhof et al., 2012; Levy and Glimcher, 2012; Bartra et al., 2013; Clithero and Rangel, 2014; Chase et al., 2015) have consistently shown that vmPFC was associated with valuation of rewards. In light of this literature, our results highlight that not only external rewards such as food or money, but also intrinsic rewards such as favorable beliefs, recruit the same valuation system. Moreover, the vmPFC was shown to automatically encode the value of objects (faces, houses, and paintings), independently of the explicit task instruction (Lebreton et al., 2009). In line with this automatic valuation and studies demonstrating unconscious motivational influences (Pessiglione et al., 2007), debriefing in our study revealed that subjects were unaware of their valence-dependent updating indicating that the valuation of belief updates need not require a voluntary process.

Extending previous work (Kuzmanovic et al., 2016a), the tracking of valence by vmPFC could be uniquely attributed to update sizes, above and beyond the influence of estimation errors, personal relevance of errors, actual base rates or final risk estimates. In other words, valence of updating depended on *improvement or worsening* of final beliefs *relative to* initial beliefs, irrespective of the beliefs or the new information per se. Moreover, we hypothesized that the valence-tracking effect in the vmPFC would be more pronounced in subjects with greater optimism bias, because belief formation should be biased by the desire to make favorable updates only if favorable updates also have a positive value. While

836 favorable future outlooks are likely to be experienced as pleasant, the sensitivity to the value
837 of such prospects may differ among individuals, dependent on their current state or their
838 personality. Indeed, subjects with a stronger optimism bias exhibited a greater valence-
839 tracking effect in the vmPFC, confirming that the vmPFC activity is sensitive to the subjective
840 value of stimuli (Grabenhorst and Rolls, 2011; Winecoff et al., 2013).

841 But what was the mechanism underlying this valence-dependent recruitment of the
842 vmPFC that was able to influence ongoing belief formation? One possibility is that in the
843 context of favorable (relative to unfavorable) updating, the vmPFC amplified incoming
844 signals and further influenced other prefrontal regions. Alternatively, vmPFC may be
845 influenced by other prefrontal regions, whose activity have already been modulated in a
846 valence-dependent manner. We tested these competing hypotheses by comparing dynamic
847 causal models comprising regions differentially recruited during favorable and unfavorable
848 updating. The models consisted of three nodes with distinct functional signatures: the dlPFC
849 represented an "update processing" node that received the exogenous input, the vmPFC was
850 included as the "valuation" node, while the dmPFC represented a "cognitive" node. The
851 dlPFC was a part of an extended network that was generally involved in updating (both
852 favorable *and* unfavorable). The same region was also engaged in tracking errors in base rate
853 estimates, while the strength of this tracking was predictive of individual learning rates. In
854 contrast, both vmPFC and dmPFC showed greater activity for favorable than unfavorable
855 updating. Having demonstrated that the vmPFC tracked the valence of belief updating in a
856 strictly-controlled task-related manner, we use the label "cognitive" node for the dmPFC to
857 distinguish it from the valuation-related vmPFC. While we cannot specify the exact kind of
858 cognitive processing associated with the dmPFC recruitment, recent meta-analyses indicate
859 clear functional dissociations with vmPFC being selectively associated with reward-related
860 tasks, and dmPFC being preferentially involved in cognitive processes such as social
861 inferences and perspective taking (Bzdok et al., 2013; de la Vega et al., 2016). In the context

862 of reconsidering one's own risk with respect to that of others, social inferences and
863 perspective taking seem highly plausible, and their valence-guided use may provide the means
864 of arriving at a particular, preferred conclusion (Kunda, 1990; Shepperd et al., 2002).
865 The Bayesian model comparison identified an optimum dynamic causal model, that
866 had a reciprocal information flow from dlPFC via vmPFC to dmPFC, with a valence-
867 dependent modulation of the coupling from dlPFC to vmPFC. This shows that particularly the
868 vmPFC filtered the incoming signal in a valence-dependent manner and influenced the
869 dmPFC accordingly. Importantly, both of these circuit features predicted individual
870 magnitudes of the optimism bias. Subjects with a stronger optimism bias showed a greater
871 increase in the dlPFC-vmPFC coupling during favorable (relative to unfavorable) updating.
872 Moreover, biased belief updating was greater, the stronger the transmission of this valence-
873 dependent signal from vmPFC to dmPFC was. Thus, the magnitude of the influence of
874 valuation on ongoing cognitive processing, mediated by the coupling from vmPFC to dmPFC,
875 predicted how much participants were biased towards more favorable updates. This finding
876 complies with previous studies on functional connectivity, where the increase in connection
877 between vmPFC ([3 51 -16]) and dmPFC ([-15 56 37]) predicted greater context-initiated
878 reevaluation of choice options across subjects (Rudorf and Hare, 2014).
879 Previous studies on effective connectivity identified the vmPFC as a *target* of the
880 directional influence of other regions. The coupling from hippocampus to vmPFC was
881 increased when people chose better remembered options (Gluth et al., 2015). Furthermore, the
882 coupling from dlPFC to vmPFC was increased during decisions to resist tempting short-term
883 rewards and to choose greater, but delayed rewards instead, and this effect was predictive of
884 between-subject differences in delay discounting (Hare et al., 2014). These and our findings
885 share the general idea of dynamic reciprocal influences between the valuation system and
886 other cognitive systems. However, our study is the first to show the opposite direction of

887 influence, namely the influence of the vmPFC on the dmPFC that mediates value-guided
888 belief formation.

889 In the resulting mechanistic model of belief formation, the valuation of ongoing
890 conclusions influences further cognitive processing, which in turn determines the final belief.
891 Thus, our results provide novel evidence for the notion that motivation to maximize pleasant
892 beliefs reinforces those cognitive processes that are most likely to yield favorable
893 perspectives. Leaving no possibility of reinterpretation of the observed effects in entirely
894 valence-independent terms, we substantially contribute to resolving the still persisting “hot
895 versus cold cognition” controversy (Kunda, 1990). As soon as we have a preference for one
896 conclusion over another, we may be in danger of automatically adjusting the knowledge we
897 recall and the inferential rules we apply in such a way as to support the preferred conclusion.
898 This bias in our reasoning has far-reaching implications for diverse decisions we make in our
899 everyday life, be it in a private or in professional contexts. While it can serve to protect from
900 discouraging and gloomy beliefs, it may also promote risk underestimations and
901 discriminating judgments.

902 **Author Contributions**

903 Conceptualization, Methodology, Software, Validation, Investigation, B.K.; Statistical

904 Analysis, B.K. and L.R.; Writing, B.K., L.R., and M.T.; Supervision, M.T.

905 **References**

906

907 Andersson JLR, Skare S, Ashburner J (2003) How to correct susceptibility distortions in spin-
908 echo echo-planar images: application to diffusion tensor imaging. Neuroimage
909 20:870-888.
910 Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-
911 analysis of BOLD fMRI experiments examining neural correlates of subjective value.
912 Neuroimage 76:412-427.
913 Bastin J, Deman P, David O, Gueguen M, Benis D, Minotti L, Hoffman D, Combrisson E,
914 Kujala J, Perrone-Bertolotti M, Kahane P, Lachaux JP, Jerbi K (2017) Direct
915 Recordings from Human Anterior Insula Reveal its Leading Role within the Error-
916 Monitoring Network. Cereb Cortex 27:1545-1557.
917 Bzdok D, Langner R, Schilbach L, Engemann DA, Laird AR, Fox PT, Eickhoff SB (2013)
918 Segregation of the human medial prefrontal cortex in social cognition. Front Hum
919 Neurosci 7:232.
920 Camille N, Griffiths CA, Vo K, Fellows LK, Kable JW (2011) Ventromedial frontal lobe
921 damage disrupts value maximization in humans. J Neurosci 31:7527-7532.
922 Chase HW, Kumar P, Eickhoff SB, Dombrovski AY (2015) Reinforcement learning models
923 and their neural correlates: An activation likelihood estimation meta-analysis. Cogn
924 Affect Behav Neurosci 15:435-459.
925 Clithero JA, Rangel A (2014) Informatic parcellation of the network involved in the
926 computation of subjective value. Soc Cogn Affect Neurosci 9:1289-1302.
927 Cumming G (2014) The new statistics: why and how. Psychol Sci 25:7-29.
928 Daunizeau J, Adam V, Rigoux L (2014) VBA: a probabilistic treatment of nonlinear models
929 for neurobiological and behavioural data. PLoS Comput Biol 10:e1003441.

930 de la Vega A, Chang LJ, Banich MT, Wager TD, Yarkoni T (2016) Large-Scale Meta-
931 Analysis of Human Medial Frontal Cortex Reveals Tripartite Functional Organization.
932 Journal of Neuroscience 36:6553-6562.
933 Diekhof EK, Kaps L, Falkai P, Gruber O (2012) The role of the human ventral striatum and
934 the medial orbitofrontal cortex in the representation of reward magnitude - an
935 activation likelihood estimation meta-analysis of neuroimaging studies of passive
936 reward expectancy and outcome processing. Neuropsychologia 50:1252-1266.
937 Eil D, Rao JM (2011) The Good News-Bad News Effect: Asymmetric Processing of
938 Objective Information about Yourself. Am Econ J-Microecon 3:114-138.
939 Eriksson J, Vogel EK, Lansner A, Bergstrom F, Nyberg L (2015) Neurocognitive
940 Architecture of Working Memory. Neuron 88:33-46.
941 Fouragnan E, Queirazza F, Retzler C, Mullinger KJ, Philiastides MG (2017) Spatiotemporal
942 neural characterization of prediction error valence and surprise during reward learning
943 in humans. Sci Rep 7:4762.
944 Friston KJ (2011) Functional and effective connectivity: a review. Brain Connect 1:13-36.
945 Garrett N, Sharot T, Faulkner P, Korn CW, Roiser JP, Dolan RJ (2014) Losing the rose tinted
946 glasses: neural substrates of unbiased belief updating in depression. Front Hum
947 Neurosci 8:639.
948 Gluth S, Sommer T, Rieskamp J, Buchel C (2015) Effective Connectivity between
949 Hippocampus and Ventromedial Prefrontal Cortex Controls Preferential Choices from
950 Memory. Neuron 86:1078-1090.
951 Grabenhorst F, Rolls ET (2011) Value, pleasure and choice in the ventral prefrontal cortex.
952 Trends Cogn Sci 15:56-67.
953 Hare TA, Hakimi S, Rangel A (2014) Activity in dlPFC and its effective connectivity to
954 vmPFC are associated with temporal discounting. Front Neurosci 8:50.

955 Hughes BL, Zaki J (2015) The neuroscience of motivated cognition. Trends Cogn Sci 19:62-
956 64.
957 Klein TA, Ullsperger M, Danielmeier C (2013) Error awareness and the insula: links to
958 neurological and psychiatric diseases. Front Hum Neurosci 7:14.
959 Kolling N, Behrens T, Wittmann MK, Rushworth M (2016) Multiple signals in anterior
960 cingulate cortex. Curr Opin Neurobiol 37:36-43.
961 Korn CW, Prehn K, Park SQ, Walter H, Heekeren HR (2012) Positively biased processing of
962 self-relevant social feedback. J Neurosci 32:16832-16844.
963 Korn CW, Sharot T, Walter H, Heekeren HR, Dolan RJ (2014) Depression is related to an
964 absence of optimistically biased belief updating about future life events. Psychological
965 Medicine 44:579-592.
966 Kunda Z (1990) The case for motivated reasoning. Psychol Bull 108:480-498.
967 Kuzmanovic B, Rigoux L (2017) Valence-Dependent Belief Updating: Computational
968 Validation. Front Psychol 8:1087.
969 Kuzmanovic B, Jefferson A, Vogeley K (2015) Self-specific optimism bias in belief updating
970 is associated with high trait optimism. Journal of Behavioral Decision Making
971 28:281–293.
972 Kuzmanovic B, Jefferson A, Vogeley K (2016a) The role of the neural reward circuitry in
973 self-referential optimistic belief updates. Neuroimage 133:151-162.
974 Kuzmanovic B, Rigoux L, Vogeley K (2016b) Brief Report: Reduced Optimism Bias in Self-
975 Referential Belief Updating in High-Functioning Autism. J Autism Dev Disord.
976 Lebreton M, Jorge S, Michel V, Thirion B, Pessiglione M (2009) An automatic valuation
977 system in the human brain: evidence from functional neuroimaging. Neuron 64:431-
978 439.

979 Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S (2017) Behavioural
980 and neural characterization of optimistic reinforcement learning. Nature Human
981 Behaviour 1.
982 Levy DJ, Glimcher PW (2012) The root of all value: a neural common currency for choice.
983 Curr Opin Neurobiol 22:1027-1038.
984 Mumford JA, Poline JB, Poldrack RA (2015) Orthogonalization of regressors in FMRI
985 models. PLoS One 10:e0126255.
986 Palminteri S, Khamassi M, Joffily M, Coricelli G (2015) Contextual modulation of value
987 signals in reward and punishment learning. Nat Commun 6:8096.
988 Palminteri S, Lefebvre G, Kilford EJ, Blakemore S-J (2016) Confirmation bias in human
989 reinforcement learning: evidence from counterfactual feedback processing. In.
990 bioRxiv, not peer-reviewed preprint.
991 Penny WD (2012) Comparing Dynamic Causal Models using AIC, BIC and Free Energy.
992 Neuroimage 59:319-330.
993 Pessiglione M, Schmidt L, Draganski B, Kalisch R, Lau H, Dolan RJ, Frith CD (2007) How
994 the brain translates money into force: A neuroimaging study of subliminal motivation.
995 Science 316:904-906.
996 Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012) Spurious but
997 systematic correlations in functional connectivity MRI networks arise from subject
998 motion. Neuroimage 59:2142-2154.
999 Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group
1000 studies - revisited. Neuroimage 84:971-985.
1001 Roese NJ, Olson JM (2007) Better, Stronger, Faster Self-Serving Judgment, Affect
1002 Regulation, and the Optimal Vigilance Hypothesis. Perspectives on Psychological
1003 Science 2:124-141.

1004 Rudman LA, Dohn MC, Fairchild K (2007) Implicit self-esteem compensation: Automatic
1005 threat defense. Journal of Personality and Social Psychology 93:798-813.
1006 Rudorf S, Hare TA (2014) Interactions between Dorsolateral and Ventromedial Prefrontal
1007 Cortex Underlie Context-Dependent Stimulus Valuation in Goal-Directed Choice.
1008 Journal of Neuroscience 34:15988-15996.
1009 Shah P, Harris AJL, Bird G, Catmur C, Hahn U (2016) A pessimistic view of optimistic belief
1010 updating. Cognitive Psychology 90:71-127.
1011 Sharot T, Garrett N (2016a) Forming Beliefs: Why Valence Matters. Trends Cogn Sci 20:25-
1012 33.
1013 Sharot T, Garrett N (2016b) Optimistic Update Bias Holds Firm: Three Tests of Robustness
1014 Following Shah et al. Consciousness and Cognition 50:12-22.
1015 Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of
1016 reality. Nat Neurosci 14:1475-1479.
1017 Sharot T, Guitart-Masip M, Korn CW, Chowdhury R, Dolan RJ (2012) How dopamine
1018 enhances an optimism bias in humans. Curr Biol 22:1477-1481.
1019 Shenhav A, Straccia MA, Cohen JD, Botvinick MM (2014) Anterior cingulate engagement in
1020 a foraging context reflects choice difficulty, not foraging value. Nat Neurosci 17:1249-
1021 1254.
1022 Shepperd JA, Carroll P, Grace J, Terry M (2002) Exploring the causes of comparative
1023 optimism. Psychologica Belgica 42:65-98.
1024 Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H,
1025 Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J,
1026 Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and
1027 structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl
1028 1:S208-219.

1029 Stephan KE, Penny WD, Moran RJ, den Ouden HE, Daunizeau J, Friston KJ (2010) Ten
1030 simple rules for dynamic causal modeling. Neuroimage 49:3099-3109.
1031 Tesser A (2000) On the confluence of self-esteem maintenance mechanisms. Personality and
1032 Social Psychology Review 4:290-299.
1033 Wessel JR, Danielmeier C, Morton JB, Ullsperger M (2012) Surprise and error: common
1034 neuronal architecture for the processing of errors and novelty. J Neurosci 32:7528-
1035 7537.
1036 Winecoff A, Clithero JA, Carter RM, Bergman SR, Wang L, Huettel SA (2013) Ventromedial
1037 prefrontal cortex encodes emotional value. J Neurosci 33:11032-11039.
1038 Xu JQ, Moeller S, Auerbach EJ, Strupp J, Smith SM, Feinberg DA, Yacoub E, Ugurbil K
1039 (2013) Evaluation of slice accelerations using multiband echo planar imaging at 3 T.
1040 Neuroimage 83:991-1001.
1041 Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale
1042 automated synthesis of human functional neuroimaging data. Nat Methods 8:665-670.
1043

Figures and Figure Legends

Figure 1. Outline and examples of experimental trials.

The figure illustrates the outline of experimental trials, including hypothetical trial examples. Each experimental trial consisted of four succeeding events: With respect to a specific adverse life event (e.g., cancer), subjects had to estimate the base rate (eBR) and their own risk (E1). They were then presented with the actual base rate (BR) and had the opportunity to estimate their own risk again (E2). After identical eBR and E1, the upper progression of the trial example shows a BR *lower* than expected indicating good news, while the lower progression shows a BR *higher* than expected indicating good news. Estimation errors (EE) corresponded to the difference between the estimated and the actual base rate, and the update corresponded to the difference between the first and the second self-risk estimate. Note that in both trial examples, the EE is 10 and the update is 8. For eBR, E1 and E2, subjects were instructed to use response buttons to adjust the displayed number to match their estimate as soon as the number font changed to green (after 2 seconds). Inter-stimulus intervals between eBR, E1, and E2, as well as inter-trial intervals after E2, were jittered and consisted of a fixation cross (not shown here).

Figure 2. Task performance and computational modeling.

A) Bars show subjects' updates, that were significantly larger after good news (GOOD) than after bad news (BAD). White dots represent simulations of updates by the 'biased' computational model that assumes asymmetric learning rates for good and bad news ('αA', two free parameters, *Alpha* and *Asymmetry*). Gray dots indicate simulated updates resulting from the 'unbiased' model that assumes identical learning rates for good and bad news ('α', one free parameter, *Alpha*). The simulated unbiased updates provide a normative benchmark for rational updating with learning rates estimated for each subject under consideration of her or his exact trial history. Error bars show standard errors. **B**) Bayesian model comparison confirmed that the biased model 'αA' best predicted subjects' updates. Model frequencies show that the majority of subjects was best described by the 'αA' model, above and beyond chance (red dashed line). Error bars show the posterior variance. **C**) Learning rates extracted from the winning model αA were significantly higher after good than bad news. Error bars show standard errors. **D)** Optimism bias ($update_{GOOD} - update_{BAD}$) and *Asymmetry* (estimated for each subject by the model 'αA') were significantly correlated (dots represent single subjects). **E-F**) Correlations between task variables separately for trials with good news (**E**) and those with bad news (**F**). * $p < .05$, **, $p < .01$.

Figure 3. Brain regions encoding errors and the valence of belief updating. **A**) When being confronted with the actual base rate, errors in base rate estimation (weighted by the personal relevance) were tracked by the anterior cingulate cortex, the inferior frontal gyrus, the anterior insula, the middle orbital gyrus and the dorsolateral prefrontal cortex (dlPFC). The line chart shows that the activity in the dlPFC (representative of all clusters) increased with decreasing error size (parametric modulation by error, negative correlation). Of all the involved regions, only in the dlPFC did the magnitude of the error tracking correlated with the general learning rate component *Alpha* (see scatter plot). Thus, subjects with a stronger error tracking in the dlPFC also more strongly adjusted their initial beliefs in response to errors. **B**) During the second risk estimation, the activity in the ventromedial prefrontal cortex (vmPFC) encoded the valence of updating, adjusted for estimation error and personal relevance. The gray box schematically illustrates the opposed valences of increasing updates after good and bad news (in this example, eBR is equal to E1). After good news, large updates are favorable as they ultimately change beliefs toward lower risk estimates and small updates are unfavorable as they let the opportunity to improve risk estimates pass by. In contrast, after bad news, large updates are unfavorable as they ultimately change beliefs toward higher risk estimates and small updates are favorable because they prevent worsening of risk estimates. Resulting valences are summarized in the table below: unfavorable (U), mid (M) and favorable (F) updates. The line chart shows that the activity in the vmPFC tracked the positive valence as it increased with increasing update sizes after good news but decreased with increasing update sizes after bad news. The scatter plot shows that subjects with a greater tracking of favorable updating in the vmPFC were also more optimistically biased in their belief updating. **A-B**): The line charts and the scatter plots were not used for statistical inference (which was carried out in parametric modulation and covariate analyses within the SPM framework); they are shown solely for illustrative purposes. **C**) After demonstrating the valence effect with the more precise parametric modulation analysis presented in B), a

simplified analysis of updating was conducted as a basis for DCM. Here, all trials were assigned to three valence categories: those with unfavorable (U), mid (M) and favorable (F) updates (adjusted for estimation error). Conjunction across these three categories revealed a distributed network involved in general updating, overlapping with the error tracking effect in the dlPFC. Comparing trials with favorable and unfavorable updates revealed the differential recruitment of the vmPFC and the dorsomedial prefrontal cortex (dmPFC) during updating. The line charts show contrast estimates in the dlPFC, vmPFC, and dmPFC, respectively.

Figure 4. Neurocircuitry mechanisms underlying optimistic belief updating. **A**) Ten different dynamic causal models varying in intrinsic connectivity and contextual modulation (unfavorable and favorable updating, U and F) were specified. The model space encompassed three brain regions involved in updating: dorsolateral prefrontal cortex (dlPFC), ventromedial prefrontal cortex (vmPFC) and dorsomedial prefrontal cortex (dmPFC). **B**) Bayesian model selection revealed that the model *m1* best explained subjects' BOLD signal, above and beyond chance (red dashed line). In this model, the coupling between dlPFC and vmPFC was differentially modulated by unfavorable and favorable updating. Hence, the vmPFC filtered the incoming information in a valence-dependent manner, and furthermore influenced the dmPFC. **C**) Connectivity parameters derived from *m1* show that the coupling between dlPFC and vmPFC tended to be weaker in the context of unfavorable relative to favorable updating. **D**) Optimism bias correlated with two parameters of *m1* (highlighted in red): Differential modulation of the dlPFC-vmPFC connection by favorable vs. unfavorable updating (F-U), and the strength of the vmPFC-dmPFC connection (vmPFC::dmPFC). Thus, subjects with a stronger optimism bias also demonstrated a greater valence-dependent filtering of incoming information by vmPFC and a greater transmission of this differential signal further to dmPFC.

Tables

Table 1. Task variables

	M (*SD*)		*p*	
Parameter	**Good News**	**Bad News**		**Source**
Number of trials	39.96 (1.45)	38.21 (2.38)	.024	
Estimated base rate (eBR)	49.74 (12.67)	45.92 (12.30)	.000	*Participants' response*
First estimate (E1)	42.64 (11.10)	37.85 (10.49)	.000	*Participants' response*
Presented base rate (BR)	36.35 (12.57)	59.76 (12.22)	.000	*Base rate algorithm*
Estimation error (EE)	13.39 (0.95)	13.84 (0.57)	.001	EE = |eBR - BR|
Second estimate (E2)	35.12 (10.86)	44.55 (11.26)	.000	*Participants' response*
Update	7.51 (2.58)	6.70 (2.20)	.045	$Update_{GOOD}$ = E1 - E2, $Update_{BAD}$ = E2 − E1
Personal relevance (PR)	0.70 (0.12)	0.69 (0.13)	.287	*for E1 < eBR:* *PR = 1 − ((eBR − E1) / (eBR − 1))* *for E1 > eBR:* *PR =1 – ((E1 − eBR) / (99 − eBR))* *for E1 = eBR:* *PR = 1*
RT eBR (s)	5.19 (0.84)	5.11 (0.80)	.192	*Participants' response*
RT E1 (s)	3.25 (0.91)	3.30 (0.90)	.461	*Participants' response*
RT E2 (s)	2.70 (0.62)	2.56 (0.61)	.018	*Participants' response*

Note. All measures (except for number of trials) were recorded or computed for each trial, and were then averaged, separately for the conditions GOOD and BAD, and separately for each participant. Positive update values indicated updates towards the BR, and negative values updates away from the BR (< 3% of the trials). PR: 1 indicates equal risk perception for the average and oneself; 0 indicates maximally different risk perception for the average and oneself; note that PR corresponds to 'relative personal knowledge' in Kuzmanovic et al., 2017. RT, reaction time. *p* values refer to paired two-tailed paired *t*-tests with $n = 24$.

Table 2. Error coding during base rate presentation and its relation to the learning rate component *Alpha*

	cluster size	peak $p_{FWE\text{-}corr}$		size		
1) Parametric modulation of BR_{GOOD} and BR_{BAD} by error						
a) Conjunction: PM_error_{GOOD} & PM_error_{BAD}, positive correlation						
No significant results						
b) Conjunction: PM_error_{GOOD} & PM_error_{BAD}, negative correlation						
Anterior cingulate cortex	63	0	5.18	10	34	20
Inferior frontal gyrus (p. triangularis)	45	0.001	4.09	46	44	0
Anterior insula	37	0.001	4.05	28	22	6
Middle orbital gyrus	30	0.002	4	16	50	-2
$dlPFC^{COV\_Alpha}$	32	0.004	3.81	40	40	28
c) PM_error_{GOOD} > PM_error_{BAD}						
No significant results						
d) PM_error_{BAD} > PM_error_{GOOD}						
Cerebellum	48	0.001	6.9	-18	-76	-46
Middle occipital gyrus	83	0.005	6.17	-34	-84	28
Superior parietal lobule	47	0.006	6.15	24	-56	48
Conjunction: PM_error_{BAD}, positive correlation & PM_error_{GOOD}, negative correlation						
Inferior occipital gyrus	22	0.001	4.16	40	-84	-10
2) Covariate analysis of error coding with *Alpha* (masked with contrast 1b)						
dlPFC	12	.008	4.46	40	38	30

Note. Error = EE * PR, based on the computational modeling of task performance. For significant differences between PM_error_{GOOD} and PM_error_{BAD} we report global conjunction results to clarify whether the difference relates to different magnitudes of the same modulation effect (e.g., the positive correlation between BOLD and error was stronger in BAD than in GOOD), or to modulation effects of opposite direction (e.g., the correlation between BOLD and error was positive in BAD, but negative in GOOD). $^{COV\_Alpha}$ indicates that in this cluster the magnitude of the error tracking correlated with the learning rate component *Alpha* across subjects (Covariate analysis). dlPFC, dorsolateral Prefrontal cortex. Peak coordinates refer to the MNI space.

1146 **Table 3**. Activity during second estimation that was modulated by update

	cluster size	peak $p_{FWE-corr}$	*T*	x	y	z
1) Parametric modulation of $E2_{GOOD}$ and $E2_{BAD}$ by update						
a) PM_update_{GOOD} > PM_update_{BAD}						
vmPFC	49	0	5.63	-12	44	-16
		0.01	5.40	-6	44	-20
Conjunction: PM_update_{GOOD}, positive correlation & PM_update_{BAD}, negative correlation						
vmPFC	44	0	3.67	-8	46	-18
		0.02	3.35	-10	54	-12
b) PM_update_{BAD} > PM_update_{GOOD}						
No significant results						
c) Conjunction: PM_update_{GOOD} & PM_update_{BAD}, positive correlation						
No significant results						
d) Conjunction: PM_update_{GOOD} & PM_update_{BAD}, negative correlation						
Fusiform gyrus (V4)	816	0	7.33	28	-72	-8
Lingual gyrus (V1)		0	4.59	6	-72	2
Lingual gyrus (V3)	736	0	6.70	-10	-86	-6
Superior occipital gyrus (V3)		0	4.11	-16	-88	18
Superior occipital gyrus	231	0	5.17	22	-80	20
Precentral gyrus	494	0	4.54	-32	-18	64
Postcentral gyrus		0	4.34	-42	-26	54
Fusiform gyrus	25	0.01	3.53	24	-46	-14
2) Covariate analysis of valence coding with optimism bias						
a) Masked with contrast 1a, conjunction						
vmPFC	39	0	15.59	-6	50	-18
b) Whole brain						
vmPFC	14	.011	7.28	-2	48	-18
3) Three categories of E2: unfavorable, mid, favorable						
a) $E2_{favorable}$ > $E2_{unfavorable}$						
$vmPFC^{DCM}$	27	0	5.84	-2	46	-22
Dorsomedial prefrontal cortex	30	0.02	5.39	-16	44	40

b) $E2_{unfavorable} > E2_{favorable}$

No significant results

c) Conjunction: all 3 categories of E2

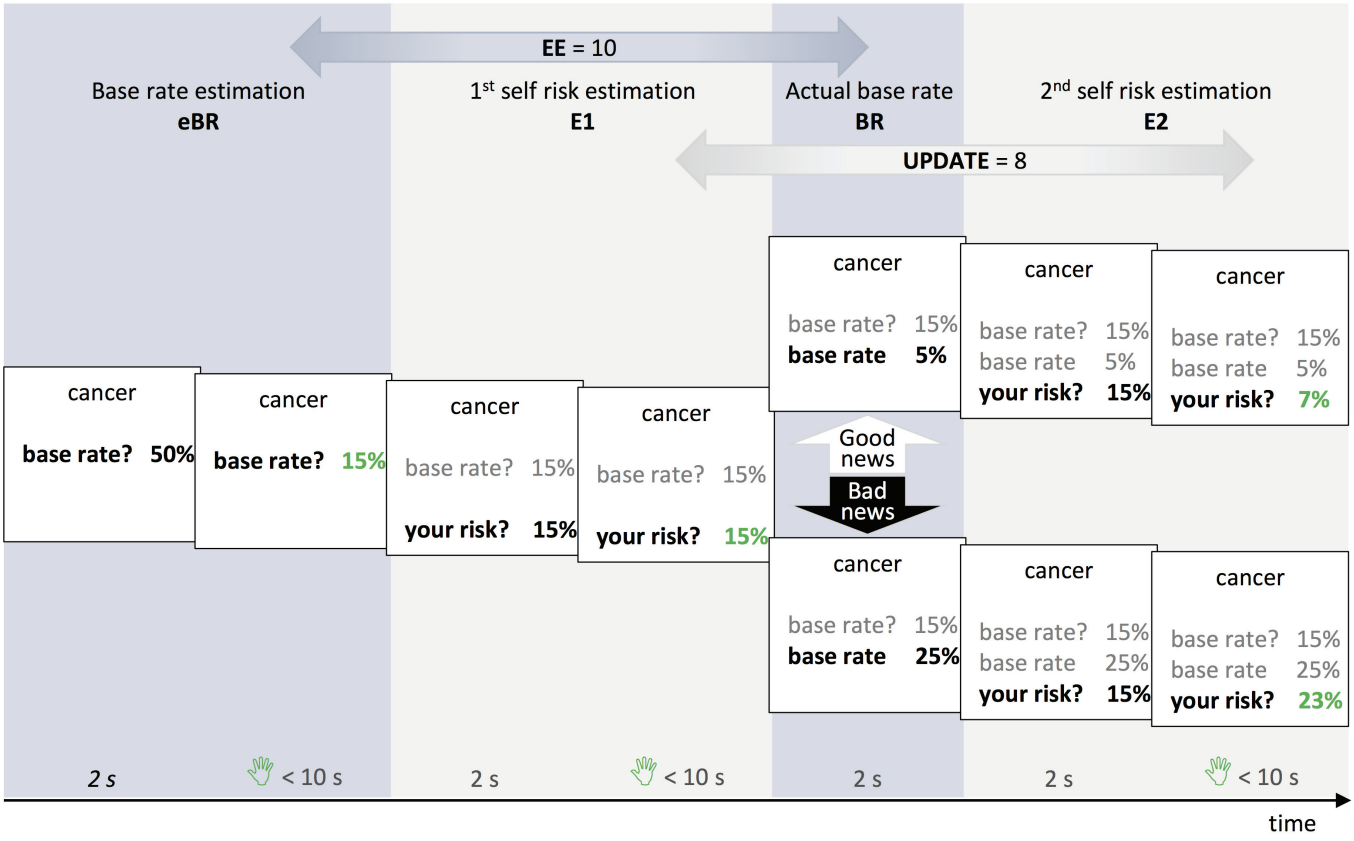
Lingual gyrus (V3)	57571	0	24	24	-86	-12
Fusiform gyrus		0	18.9	-30	-58	-14
Lingual gyrus (V4)		0	18.7	-24	-86	-14
Fusiform gyrus		0	16.4	32	-50	-18
Inferior parietal lobule		0	14.8	-44	-40	48
Inferior parietal lobule		0	12.6	50	-34	48
Thalamus		0	8.38	20	-30	-2
Middle frontal gyrus	6959	0	10.5	42	2	60
Inferior frontal gyrus (p. opercularis)		0	9.62	46	10	36
$dlPFC^{DCM}$		0	8.99	44	42	26
dlPFC	1414	0	8.46	-28	52	28
dlPFC		0	8.21	-40	30	32
Inferior frontal gyrus (p.triangularis)		0	7.44	-34	34	24
Thalamus	136	0	6.58	-8	-22	8
Posterior cingulate cortex	125	0	6.13	-2	-24	28
Precentral gyrus	23	0.01	5.65	34	-28	72

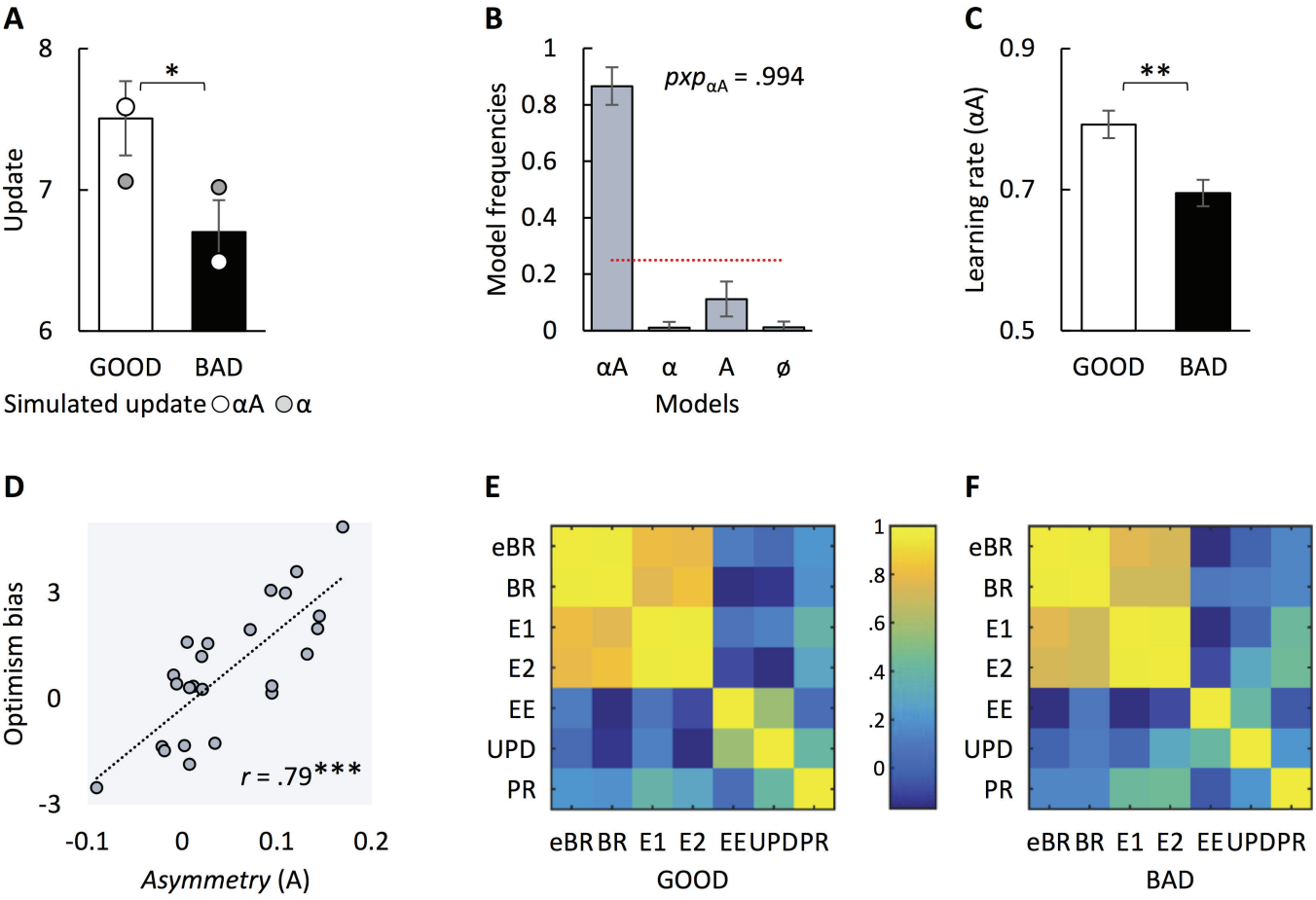
Note. For significant differences between PM_update_{GOOD} and PM_update_{BAD} we report global conjunction results to clarify whether the difference relates to different magnitudes of the same modulation effect (e.g., the positive correlation between BOLD and update was stronger in GOOD than in BAD), or to modulation effects of opposite direction (e.g., the correlation between BOLD and update was positive in GOOD, but negative in BAD). vmPFC, ventromedial prefrontal cortex; dlPFC, dorsolateral prefrontal cortex. Peak coordinates refer to the MNI space.

Table 4. DCM parameter estimates of the model *m1* and correlations with measures of optimism bias

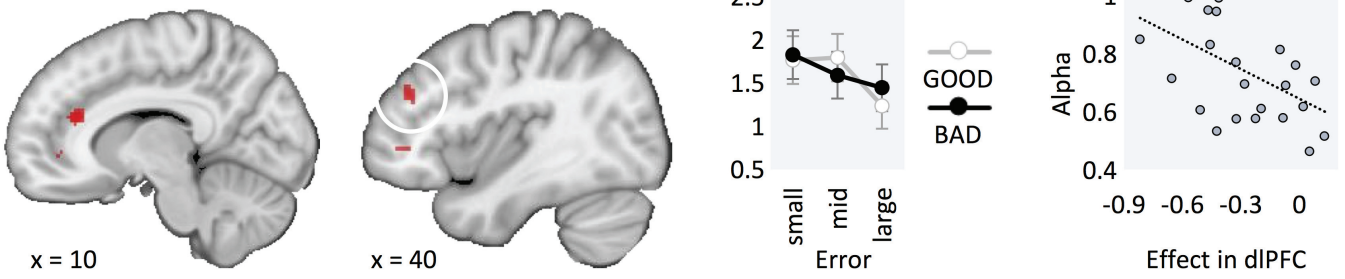
	M (*SD*)	*p*	*r*	*p*	*r*	*p*
		t-test	Optimism bias		*Asymmetry*	
Matrix A						
dl::dl	-0.01 (0.09)	0.000*	-0.19	0.361	-0.21	0.323
dl::vm	-0.06 (0.12)	0.030	-0.23	0.277	-0.05	0.819
vm::dl	0.10 (0.15)	0.000*	-0.09	0.679	-0.15	0.485
vm::vm	-0.07 (0.09)	0.000*	-0.20	0.348	-0.20	0.342
vm::dm	0.15 (0.14)	0.000*	0.47	0.020*	0.49	0.015
dm::vm	0.08 (0.124)	0.004*	0.27	0.207	0.32	0.127
dm::dm	-0.02 (0.03)	0.001*	-0.49	0.015	-0.56	0.005
Matrix B						
U on dl::vm	-0.35 (1.06)	0.123	-0.23	0.275	-0.31	0.147
F on dl::vm	0.05 (1.00)	0.822	0.49	0.015	0.36	0.088
F - U	0.39 (1.42)	0.188	0.52	0.009*	0.48	0.018
Matrix C						
U M F to dl	0.12 (0.07)	0.000	0.20	0.340	0.13	0.541

Notes. Parameter estimates in Hertz, self-connections were log-transformed. dl, dorsolateral prefrontal cortex; vm, ventromedial prefrontal cortex; dm, dorsomedial prefrontal cortex; ::, endogenous connection; U, unfavorable updating; M, mid updating; F, favorable updating. *, equivalent to $p < .05$, Bonferroni-corrected for multiple comparisons (Matrix A, t-tests, $p < .007$ corrected for 7 comparisons; *r*, optimism bias, $p < .025$ corrected for 2 comparisons with a priori hypotheses).

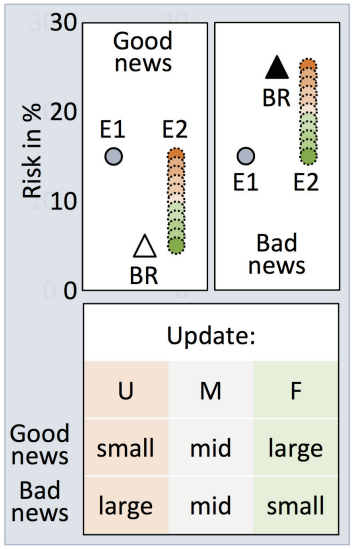


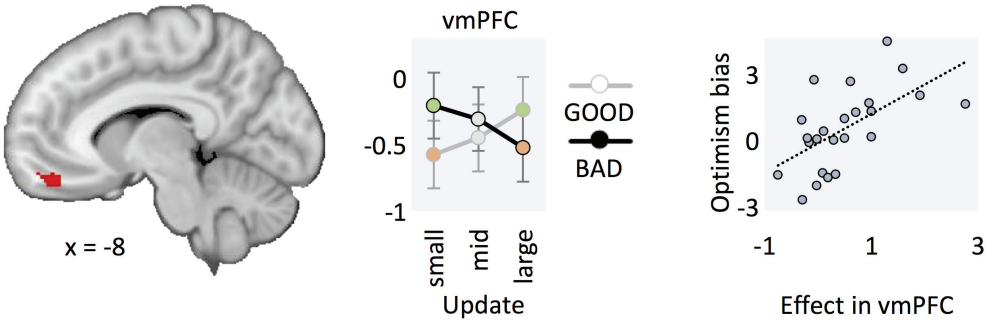


A Error tracking

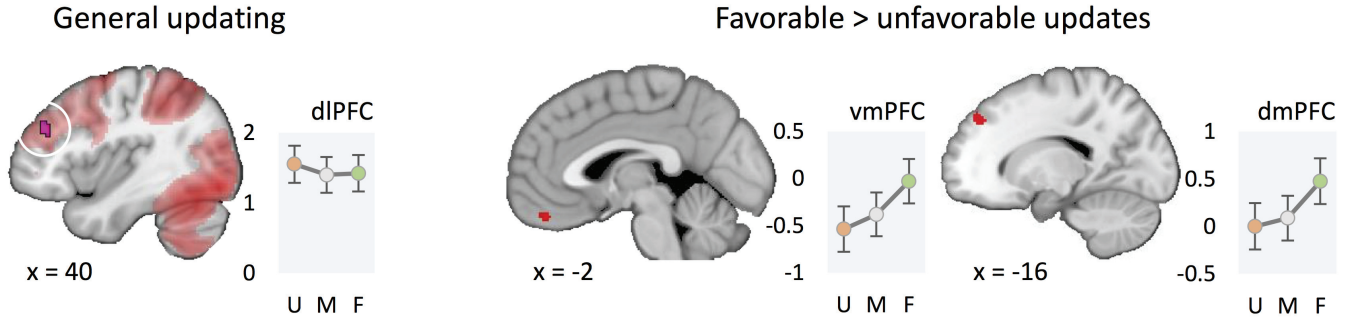


B The valence of updating





C Three categories of updates



U, Unfavorable updates; F, Favorable updates; $p_{FWE\text{-}corr.} < .05$, peak level, whole brain; Error tracking effect